University of Bath

PHD

Numerical Model Error in Data Assimilation

Jenkins, Sian

Award date:
2015

Awarding institution:
University of Bath

Link to publication

Alternative formats
If you require this document in an alternative format, please contact:
openaccess@bath.ac.uk

Copyright of this thesis rests with the author. Access is subject to the above licence, if given. If no licence is specified above, original content in this thesis is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC-ND 4.0) Licence (https://creativecommons.org/licenses/by-nc-nd/4.0/). Any third-party copyright material present remains the property of its respective owner(s) and is licensed under its existing terms.

Take down policy
If you consider content within Bath's Research Portal to be in breach of UK law, please contact: openaccess@bath.ac.uk with the details. Your claim will be investigated and, where appropriate, the item will be removed from public view as soon as possible.

Download date: 06. May. 2023

Numerical Model Error in Data Assimilation

submitted by

Siân E. Jenkins

for the degree of Doctor of Philosophy in Engineering and Mathematics

of the

University of Bath

Department of Electronic and Electrical Engineering

November 2014

COPYRIGHT

Attention is drawn to the fact that copyright of this thesis rests with the author. A copy of this thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with the author and that they must not copy it or use material from it except as permitted by law or with the consent of the author.

This thesis may be made available for consultation within the University Library and may be photocopied or lent to other libraries for the purposes of consultation with effect from .....

Signed on behalf of the Faculty of Engineering.....

ACKNOWLEDGEMENTS

First of all, I would like to thank my supervisors Dr. Nathan Smith, Prof. Chris Budd and Dr. Melina Freitag for all their guidance and help over the past four years and for their patience with my verbose style of writing. I would especially like to thank Nathan for all the time and energy he put into understanding my work and for trusting and allowing me to work in an office in the Department of Mathematical Sciences. I would also like to thank the University of Bath for awarding me a University Research Studentship to support my studies.

Next I would like to thank Dr. Adrian Hill and Dr. Euan Spence for their discussions on the CFL number and 2D integration by parts, respectively. I would also like to thank Dr. Adrian Hill and Dr. Amos Lawless for agreeing to be my examiners and for taking the time to read this lengthy thesis.

I dedicate this thesis to my family - Mum, Dad and Ceri. Without you there would have been no thesis. You were my rocks in my turbulent first year and always there to listen and provide me with a hug when I needed one.

The support and energy of the students in the Maths department has been amazing. Thanks to Ray, Andrea, Aretha, Lisa, Karin, Amy, Jen, Jack, Steve, Katy, Sam, Doug, Elvijs, Alex. W and Amine to name a few. Also to the crew in 4W1.15: Andrew, Mason, (Fake) Finn, James. C, Alex. C, Steven, Elizabeth and especially Hannah. I am so glad I completed my PhD with you all. The numerous social events have helped me through my PhD: Drinks in the Parade, Christmas and Summer graduation parties, DitPoT to Avoncliff, board games and cake. I never quite got the hang of drinking but your quiet encouragement has meant that I am starting to enjoy it.

Thanks also to Alex. W and Charlie - I will never read board game instructions again without a funny accent (and perhaps some wine)!

Finally, I would like to thank Eugene Duffy and Chloe Lock for all their generosity and support whilst I lived with them. I was always happy to come home because I knew you were both there, even if I was never sure what else I might find. I loved living with Barney (dog) and the girls (the chickens). Thank you for making me feel apart of a family whilst I was away from mine and for taking me to Penzance for a long weekend when I needed it.

SUMMARY

In this thesis, we produce a *rigorous* and *quantitative analysis* of the errors introduced by finite difference schemes into *strong constraint 4D-Variational (4D-Var) data assimilation*. Strong constraint 4D-Var data assimilation is a method that solves a particular kind of inverse problem; given a set of observations and a numerical model for a physical system together with a priori information on the initial condition, estimate an improved initial condition for the numerical model, known as the analysis vector. This method has many forms of error affecting the accuracy of the analysis vector, and is derived under the assumption that the numerical model is perfect, when in reality this is not true. Therefore it is important to assess whether this assumption is realistic and if not, how the method should be modified to account for model error. Here we analyse how the errors introduced by finite difference schemes used as the numerical model, affect the accuracy of the analysis vector.

Initially the 1D linear advection equation is considered as our physical system. All forms of error, other than those introduced by finite difference schemes, are initially removed. The error introduced by 'representative schemes' is considered in terms of *numerical dissipation* and *numerical dispersion*. A spectral approach is successfully implemented to analyse the impact on the analysis vector, examining the effects on unresolvable wavenumber components and the l_2-norm of the error. Subsequently, a similar also successful analysis is conducted when observation errors are re-introduced to the problem. We then explore how the results can be extended to weak constraint 4D-Var.

The 2D linear advection equation is then considered as our physical system, demonstrating how the results from the 1D problem extend to 2D. The linearised shallow water equations extend the problem further, highlighting the difficulties associated with analysing a coupled system of PDEs.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

This thesis presents a *rigorous* and *quantitative analysis* of the effects of numerical model error, introduced by finite difference approximations, on the results of *strong constraint 4D-Variational (4D-Var) data assimilation*. Strong constraint 4D-Var solves a particular inverse problem; given observations and a numerical model for a physical system together with a priori information on the initial condition for the numerical model, strong constraint 4D-Var calculates an improved estimate for the initial condition, known as the *analysis vector*. The analysis vector can then be used as the initial condition for the numerical model, to create a prediction for the physical system. Strong constraint 4D-Var is one of many data assimilation methods. Each data assimilation method has been designed to solve a similar inverse problem. The analysis vector is identified by minimising the strong constraint 4D-Var cost function with respect to the initial condition for the numerical model. The weighted least squares formulation of the cost function aims to minimise the effects of errors associated with variables in the cost function, such as observation errors. However model errors are one form of error whose affects are not accounted for as the derivation of the method assumes the model to be perfect [2].

Model errors can arise from several difference sources such as inaccurate model equations and numerical model errors that arise due to errors in numerical implementation. This thesis focuses on numerical model errors introduced by solving the model equations using finite difference schemes. The finite difference schemes introduce numerical model errors through the approximation of derivatives by finite differences. The impact of model errors is usually small in comparison to others, such as observation errors [3]. However, it could give rise to artifacts in the analysis vector and the forecast made using it. Therefore it is important we understand its impact.

In recent years, the method of strong constraint 4D-Var data assimilation has become widely used in operational weather centres for numerical weather prediction (NWP), due to its efficient implementation through the method of incremental 4D-

Var [4]. This is an iterative method involving inner and outer loop processes to attain the analysis vector [5]. As this method is becoming more common place, it is important that we understand whether the assumption of a perfect model is reasonable. If this assumption is not appropriate, then a way of correcting for the effects of model error is required. In order to understand and possibly achieve this goal, the effects need to be quantified and analysed.

Initially, the physical system chosen to investigate the effects of numerical model error on 4D-Var, is the linear advection equation. Many of the physical systems considered in numerical weather prediction involve "wave-like flow" [6]. The linear advection equation is one such model. It is a linear, hyperbolic PDE [6]. Here we choose it as a representative model (and a prototype) for more complex advective processes of interest in NWP. This system provides a numerical challenge despite looking deceptively simple [7]. Kreiss [8] developed his theory of numerical stability using the linear advection equation, revealing just how challenging the equation can be. The linear advection equation also has the property that some complex linear hyperbolic systems can be decomposed into the superposition of solutions to several linear advection problems [6]. The 1D linearised shallow water equations is one such system under certain assumptions [9]. This is why we choose to investigate the linearised shallow water equations once our analysis of the linear advection equation is complete. However to prevent the system of equations from decoupling into a system of linear advection equations, the system is considered in 2D together with Coriolis acceleration.

Another reason to begin with a linear problem is the relevance of linearised numerical models to incremental 4D-Var. Most of the physical systems considered for use in strong constraint 4D-Var have non-linear models. Consequently, their numerical models are non-linear and the cost function is no longer quadratic. This leads to issues with the computational cost, implementation time and difficulties with isolating the global minimum of the problem [10]. Incremental 4D-Var was developed to reduce these problems by using a *tangent linear model (TLM)* assumption, requiring the linearisation of the non-linear numerical model, about the current model state [5, 11, 12]. The linearisation requirement of incremental 4D-Var, makes investigating linear models essential, as well as a good first step to begin our analysis.

The linear advection equation also has the added bonus of possessing an analytical solution. Therefore the results of any finite difference scheme can be compared to the analytical solution [13]. This gives rise to the question 'how do you determine the accuracy of a finite difference scheme?'. The answer to this question depends on the desired application of the scheme. The l_2-norm of the error can be used to quantify the error, but it does not provide any information on the ability of the scheme to propagate the unresolvable wavenumber components of the initial condition. Unresolvable wavenumber components arise due to the finite grid the finite difference schemes are defined on. Durran [6] advocates the use of a *"spike test"*, where the ability of the scheme to propagate an initial condition which is flat apart from a large spike, is

tested. This test reveals the ability of a scheme to propagate unresolvable wavenumber components, which make up discontinuous initial conditions [6]. Applications such as tsunami detection require an accurate method for determining the presence of a tsunami and predicting its motion. Dam break problems and tsunami waves can contain shock profiles [9] made up of lots of high wavenumber components. As the analysis vector resulting from strong constraint 4D-Var is used to predict the motion of such physical systems, it is important that we assess the impact of numerical model error on the ability of strong constraint 4D-Var to reconstruct these profiles. Hence we investigate the effects of numerical model error on strong constraint 4D-Var data assimilation, through both the l_2-norm of the error as it is "closely related to conserved physical quantities" [6] and through the use of a shock profile as the true initial condition.

We analyse the effects of numerical model error through the *numerically dissipative* and *numerically dispersive* properties of the considered finite difference schemes. The effects of numerical dissipation and dispersion on the results from a finite difference scheme are well understood for systems constructed from a single PDE [14, 15]. The effects on an inverse problem, such as strong constraint 4D-Var data assimilation, are not well understood. A spectral approach is taken is taken in our analysis, similar to the analysis of the results from finite difference schemes.

In Chapter 2 we present some of the theory and variables associated with data assimilation, before examining the method of strong constraint 4D-Var data assimilation in detail. We then consider the method of incremental 4D-Var used to efficiently numerically implement strong constraint 4D-Var. A review of some of the literature surrounding numerical model error in data assimilation is also presented, including weak constraint 4D-Var data assimilation. The Chapter concludes by stating the strong constraint 4D-Var data assimilation problem to be considered for various physical systems, in this thesis.

Chapter 3 considers the impact of finite difference approximations on strong constraint 4D-Var data assimilation, by choosing the 1D linear advection equation together with circulant boundary conditions and initial condition, as our physical system. We begin our analysis by discussing the theory surrounding 1D Fourier series, including their convergence properties and discrete Fourier series. This allows the concepts of numerical dissipation and dispersion to be introduced. We investigate our problem through three 'representative schemes' used to solve our physical system; the Upwind, Preissman Box and Lax-Wendroff finite difference schemes. These schemes are chosen due to their numerically dissipative and dispersive properties. We develop a numerically non-dissipative and non-dispersive scheme, with respect to the resolvable wavenumber components of the numerical solution, the MNIMC scheme. This scheme allows perfect observations to be constructed in a convenient fashion for use algebraically in our analysis and to define metrics for measuring the numerically dissipative and dispersive properties of a scheme. The aliasing error in the MNIMC scheme is found to possess a shifted periodic nature. Using a spectral approach, we are able to construct the analysis

vector from the matrices implementing one of our considered schemes and the MNIMC scheme, along with an aliasing correction term. This allows the effects of numerical dissipation and/or dispersion on the analysis vector, to be viewed directly. Hence the quality of our analysis vector can be assessed, for true initial conditions with 'shock profiles', similarly to Durran's "spike test" [6]. We perform this analysis in the absence of all forms of error in the data assimilation problem, apart from numerical model error introduced by finite difference schemes.

Chapter 4 continues our analysis of the problem from Chapter 3. A bound on the l_2-norm of the error in the analysis vector is created through spectral methods. This bound is analysed to determine its suitability for representing the behaviour of the error in the analysis vector with respect to the number of discretisation points when considering full sets of observations, the number of sets of observations in the assimilation window, the numerically dissipative and dispersive properties of the finite difference schemes and the smoothness of the true initial condition. This is achieved by numerically generating the l_2-norm of the error in the analysis vector for various regularity initial conditions and finite difference schemes, allowing the behaviour of the bound and the numerical results to be compared. Having completed this analysis, observation errors are re-introduced to the problem and a similar analysis is performed. The contribution from observation errors to the analysis vector is analysed to determine if correlations are introduced by strong constraint 4D-Var, which could possibly lead to artifacts in the analysis vector. We then go on to relate our work to developing the deterministic model error operator for the weak constraint 4D-Var data assimilation problem. The results of Chapters 3 and 4 have been submitted for publication in the Journal of Computational and Applied Mathematics.

In Chapter 5, we extend our problem to the 2D linear advection equation together with circulant boundary conditions and initial condition. We again consider the impact of finite difference approximations from finite difference schemes on the accuracy of the analysis vector. A similar analysis to that of Chapters 3 and 4 is conducted, in the absence of all other forms of error. Numerical model error can be considered in the form of numerical dissipation and dispersion as before. A bound is constructed for the l_2-norm of the error in the analysis vector, for true initial conditions that are multiplicatively separable. In order to construct this bound, a bound on the 2D Fourier coefficients and a bound on the error in the coefficient identified by the 2D DFT, in comparison to the coefficient of the 2D Fourier series for the same resolvable wavenumber component, are derived for multiplicatively separable functions. We discuss the challenges associated with forming a bound on the 2D Fourier coefficients of functions which are not multiplicatively separable. Numerical results for the l_2-norm of the error in the analysis vector, due to finite difference approximations, are presented.

Chapter 6 considers the same strong constraint 4D-Var problem with the 2D linearised shallow water equations, together with circulant boundary conditions and initial condition, as our physical system. We begin by investigating the building blocks re-

quired to explore the effects of numerical model error, introduced by finite difference schemes, on the analysis vector. Whilst exploring the finite difference schemes for solving the 2D linearised shallow water problem, we discover the difficulties associated with defining numerical dissipation and dispersion for a coupled system of PDEs. We discuss how the matrix polar decomposition may be a possible way to define such quantities. The MNIMC scheme for the problem is then constructed in the hope that it could be used to create perfect observations for our numerical experiments. As a part of this analysis, we also investigate the properties of the aliasing error introduced by the MNIMC scheme, for solving the 2D linearised shallow water problem. Through this analysis, we discuss the difficulties involved in numerically generating perfect observations of the 2D linearised shallow water problem, for use in numerical strong constraint 4D-Var experiments.

Chapter 7 summarises the conclusions made from the research presented in this thesis. The novelty in this thesis lies in:

- Quantifying and qualifying the error in the analysis vector due to numerical model error introduced by finite difference approximations in the forward model.
- Investigating the effects of numerically dissipative and dispersive schemes on the contribution to the analysis vector from observation errors.
- Identifying a deterministic model error operator and the number of random variables required to augment the numerical model for use in weak constraint 4D-Var data assimilation, of the 1D linear advection problem.
- The development of numerically dissipative and dispersive metrics for measuring the effects of these errors on the numerical solution generated by finite different schemes, for solving the considered linear advection problems.
- The development of a finite difference scheme that is numerically non-dissipative and non-dispersive with respect to the resolvable wavenumber components of the numerical solution.
- Attempting to expand the definition of numerical dissipation and dispersion to linear systems of PDEs.

In particular, through the following results:

- Numerically non-dissipative and dispersive finite difference schemes for solving the 1D linear advection problem, introduce destructive interference between wavenumber components, leading to a loss of information in the analysis vector.
- There exists a critical number of discretisation points when considering full sets of observations, where the effects of finite difference scheme errors and observation errors on the accuracy of the analysis vector are minimised, for the 1D linear advection problem.

- The contribution to the analysis vector from white noise observation errors can become correlated due to the effects of numerical dissipation in the resolvable wavenumber components of the scheme, possibly leading to artifacts in the analysis vector.
- Increasing the number of sets of observations in the assimilation window does not necessarily alter the contribution of high real resolvable wavenumber components to the analysis vector.
- Numerical dissipation and/or dispersion may help to reduce the impact of other forms of error on the accuracy of the analysis vector.

The analysis in this thesis was far more technical than expected. As a result there are substantial appendices included after the Conclusions in Chapter 7. The main results lie in Chapters 3-6. We now begin our analysis with Chapter 2, presenting some of the theory behind data assimilation and discussing some of the topics discussed in this Introduction, in more detail.

CHAPTER 2

Data Assimilation

In this chapter, we explore the literature surrounding data assimilation and the effects of model error on its results. We begin by setting out the problem that data assimilation methods aim to solve, along with the errors associated with it. We then examine variational data assimilation methods, including 3D-Variational (3D-Var) and 4D-Variational (4D-Var) data assimilation techniques. Next we describe the operational implementation of 4D-Var data assimilation through *incremental 4D-Var data assimilation*, used in operational weather forecasting centres. The chapter concludes by discussing previous research that has been conducted into the effects of model error on the results of data assimilation.

2.1 The data assimilation problem

Data assimilation methods are designed to solve a particular kind of inverse problem, given by the following.

Given observations and a numerical model for a physical system, estimate the true state of the physical system.

This problem arises in many fields, such as oceanography, hydrology and numerical weather prediction (NWP) [16], as well as others that can be found in [16, 17]. Many methods have been developed for solving this inverse problem, such as Optimal Interpolation (OI), Best Linear Unbiased Estimator (BLUE), the Kalman Filter and Variational methods [10, 18, 19]. Under certain assumptions, some of these methods become equivalent [18]. We are particularly interested in the variational techniques available for solving this problem and will go into more detail about these in the following sections. Details on the other methods can be found in [10, 11, 16, 18, 19, 20, 21].

The idea behind data assimilation is that better estimates of the true state of the physical system can be made by incorporating "time distributed observations and a dynamical model", rather than relying on the interpolation of observations, creating a model which is consistent with the physical system [10]. The data assimilation process makes use of the provided observations and the numerical model to make a " 'best' estimate" [2] for the true state of the physical system. The term " 'best' estimate" [2] refers to the fact that the type of estimate used to solve the data assimilation problem, is problem dependent. For example, estimates for the mean or mode of a desired quantity could be made [2, 10]. Variational methods produce an estimate for the mode of the *probability density function* (pdf) of the initial condition for the numerical model, producing a maximum likelihood estimate for the initial condition [2]. A minimum variance estimate for the initial condition is produced by estimating the mean of the pdf of the initial condition [2].

The method chosen to solve a particular instance of the data assimilation problem, also depends upon factors such as the number of observations in comparison to the number of state variables, the statistical properties of the errors associated with the variables and the size of the considered problem. The methods can be derived from many different fields of study, such as *control theory*, *Bayesian statistics* and *maximum likelihood estimation*, to name a few [10, 22].

The numerical model for the physical system is often referred to as the *forward model* for the problem [2]. It solves the model equations across a finite grid, forward in time. The accuracy of the results from the forward model, is dependent on the resolution of the grid. Once the data assimilation problem has been solved, in some applications, it is desirable to use the result in the forward model to generate a forecast for the physical system. This is the case in NWP. Variational techniques estimate an improved initial condition for the forward model, to allow a forecast to be generated past the time of the last considered observation.

The physical systems where data assimilation is applied, tend to be chaotic, non-linear systems [10]. Solving such systems numerically is challenging due to the computational costs involved in solving non-linear systems. Their chaotic nature also means that small errors in the state of the system, estimated through data assimilation, can become much larger as the numerical model calculates a forecast for the system. The Lorenz system is an example of a highly chaotic system. The observations, the numerical model and many other aspects of the data assimilation process, all contain errors. Therefore the solution obtained through data assimilation, will not be perfect. As a result, we would expect the forecast to "diverge from the truth, after a finite time" [10]. In order to combat this, new observations are taken of the physical system, periodically. Data assimilation can then be conducted at this future point in time to re-calibrate the forecast.

In order to demonstrate the errors associated with data assimilation, NWP is used as an example application in the following paragraphs. This application demonstrates

many of the constraints that can be placed on the solution of data assimilation problems. Two important constraints are the limitations of computational resources and the time in which the problem needs to be solved. The atmosphere is a three dimensional, highly chaotic and inter-connected, dynamical system. The weather at any one point in the system, is affected by the surrounding weather in all directions. This includes the oceans and space weather systems at the boundaries of the atmosphere. Therefore in order to form the best forecast for the system, we may need to model the weather around the entire Earth, coupling it with ocean and space weather models. However modelling such a system numerically requires a lot of computational resources. This limits the resolution of the grid, affecting the accuracy of the model. The UK Met. Office have a supercomputer to perform data assimilation and generate weather forecasts using its Unified Model, which has a coupled atmospheric and ocean model [23]. These processes all need to be completed ahead of the time period the forecast is generated for. This also places a limit on the complexity and grid resolution of the model.

The global model configuration of the Unified Model is used to generate medium-range weather forecasts every six hours, using a 17km resolution, for a forecast of six days [24]. In order to reduce the time and computational resources required, some aspects of the numerical models do not need to be executed. For example, the Met. Office may not run their ocean component when making "short-range weather forecasting" using "a higher resolution atmospheric model" [23]. On areas of specific interest, finer grids are placed to provide a greater level of accuracy in forecasts [24]. The Met. Office use a "variable resolution UK model" with a 1.5km resolution over locations of "forecast interest". This is surrounded by a coarser grid with a 4km resolution at the boundaries of the model. A variable resolution grid bridges the gap between the two resolutions [24].

Observations

The observations of the physical system are made using various different techniques. In the case of NWP, observations are made of variables such as the wind speed, temperature and pressure. These need to be made across the world and at various altitudes. Observations are taken at both stationary and mobile weather stations. These include ground based stations, buoys and ships travelling across the world oceans [20]. In order to gain information about the weather at various altitudes, weather balloons and satellite data are some of the sources of information available to meteorologists. The nature of non-satellite based observations means that the density of observations is greater over highly populated regions of the Earth [25]. Using only these observations impacts the accuracy of the forecasts over less densely observed regions. The use of satellite observations has helped to increase the number of observations taken over less readily observable areas of the world.

Observations of any system contain errors. These could be due to instrument mis-

calibration or drift [18]. Attempts are sometimes made to compensate for systematic errors and biases in observations [10, 11]. The paper by Móller and Raschke [26] demonstrates the types of error that can enter into the different types of observations used in NWP. As we do not know the true state of the system, we cannot quantify the size of the errors involved. However, there are methods available for estimating information on error statistics [18].

Despite all of these observation methods, the number of observations tends to be sparse when compared to the number of grid points the numerical model is being solved across and smaller than the number of state variables the numerical model is solving for, leading to an underdetermined problem [18]. NWP typically involves solving for $\mathcal{O}(10^7)$ state variables [17], using $\mathcal{O}(10^6)$ observations [5]. This results in an ill-posed problem to be solved through data assimilation [16]. As a result, the solution techniques make use of a priori information to constrain the problem, so that it becomes well-posed. In the case of variational methods, as the aim is to estimate an improved initial condition for the numerical model, a priori information on the initial condition is used to constrain the problem, in the form of an estimated initial condition. A discussion on the ill-posedness of inverse problems can be found in [16].

The observations of the physical system are denoted by the vectors $\mathbf{y}_l \in \mathbb{R}^{m_l}$, $m_l \in \mathbb{N}$ for all $l = 0, \ldots, L \in \mathbb{N}$. The vector $\mathbf{y}_l$ contains the lth set of observations of the physical system. This includes every observation taken of any variable, at any spatial location, taken at the lth time. The observations are not necessarily taken at equally spaced points in space or time or of the same variables. The states of all variables associated with the 4D-Var cost function in (2.1), are generated at the times of the observations to allow for comparison. Any variable with subscript l, is the relevant variable at the time of the lth set of observations. The period of time the observations are taken over is known as the *assimilation window*. The cost function in Equation (2.1) uses $L + 1$ sets of observations. The errors in the observations are assumed to be unbiased, serially uncorrelated, Gaussian random variables [17].

Observations are pre-processed to remove anomalous data. This involves a "quality control check" where the data is checked for obvious errors, for example by cross-checking the type of observation with its method and location of observation [11]. Extreme observations are also rejected. These are identified by comparing an observation with those surrounding it and the forecast generated by the background estimate at that location [11]. As previously mentioned, biases and systematic errors introduced to observations can sometimes be compensated for. This is because knowledge on these forms of error can be gained from sources such as instrumentation calibration data. A discussions on compensating for model bias can be found in Lawless [11] and Dee and Da Silva [27].

The background estimate

The use of a priori information to constrain the data assimilation problem, also introduces errors into the data assimilation process. This estimated data is referred to as *background data* in NWP literature and consequently the errors in it are termed *background errors*. Data assimilation methods need to take into account these errors, along with those present in other variables associated with solving the problem.

The variable $\mathbf{x}_b \in \mathbb{R}^N$ is an estimated initial condition, providing a priori information on the initial condition for the numerical model, denoted by $\mathbf{x}_0 \in \mathbb{R}^N$. This information constrains $\mathbf{x}_0$, ensuring that the solution is well-posed, as discussed in the previous Section. Here $N \in \mathbb{N}$, denotes the dimension of the state vectors of the numerical model.

The vector $\mathbf{x}_b$ is termed the *background estimate* and is determined from previous knowledge. In the case of NWP, this could be in the form of historical weather records [20]. However, as a series of forecasts is being made, the background estimate is generally obtained from the previous forecast. As $\mathbf{x}_b$ is an estimated initial condition for the numerical model, it has errors associated with it. The errors are assumed to be Gaussian random variables and to be uncorrelated with observation errors [17]. This results in a background error covariance matrix $B \in \mathbb{R}^{N\times N}$ that is symmetric positive definite [18]. Research is being conducted into the structure of this matrix [28] as it is important that it expresses the "correlation between the errors of different variables" of the physical system accurately [10]. A practical implementation of the background error covariance matrix can be found in [29]. Some recent developments on the background error covariance matrix can be found in [30, 31].

The observation operator

Another problem associated with observations, is that they may not be observations of a variable in the numerical model. They may need to be transformed to obtain an observation for the required variable. If we are observing a variable of the numerical model, then the observation is termed a *direct observation*. However if we need to transform the variable, this is termed an *indirect observation*. The transformation could be as simple as converting a temperature from Fahrenheit to Celsius, or it could involve complex calculations. The latter is true for radio occultation data, where the Abel transform is used in the conversion of bending angle to a temperature profile [32]. Numerical implementations of these complex transforms, obtain numerical approximations to the required solution. As a result there are errors associated with these transforms.

We have already discussed that the observations of the system are sparse in comparison to the grid points where the numerical model produces a solution. This means that there may be grid points of the numerical model, where there are not any nearby observations. In this situation, the use of the numerical model helps to maintain the

consistency of the model in these areas [10, 33]. The results from some analyses have shown a greater distance between observations can increase the accuracy of strong constraint 4D-Var experiments [17], indicating that sparsity of observations may be an advantage. Observations are made at any point in space. Therefore model variables need to be interpolated to nearby observation locations for comparison. This introduces interpolation errors into the problem. These errors, along with those introduced by transforming indirect observations, are known as *representative errors* [11, 18].

The function $\mathcal{H}_l : \mathbb{R}^N \to \mathbb{R}^{m_l}$ is the lth possibly non-linear, observation operator mapping the state of the numerical model, to the state space of the lth set of observations. This function performs all the necessary transformations and interpolations discussed in the previous paragraphs. It transforms the state of the numerical model $\mathbf{x}_l$, into the same variables at the same locations as the observations in $\mathbf{y}_l$, for comparison. Therefore representative errors enter into 4D-Var data assimilation through this function. The process of constructing these operators can be quite complex, so can require a lot of effort [10].

The matrices $R_l \in \mathbb{R}^{m_l \times m_l}$ for $l = 0, \ldots, L$, are symmetric positive definite matrices, where R_l is the lth observation error covariance matrix [18]. Despite the matrices being named for observation errors, R_l is the covariance matrix for the observation errors in $\mathbf{y}_l$ and the representative errors in $\mathcal{H}_l(\cdot)$. The representative errors are also assumed to be Gaussian random variables [2], which is a good assumption for meteorological systems [10]. There will be some variables where the error statistics are not Gaussian in reality. However, many distributions can be transformed to Gaussian distributions [10], which is an advantage of this assumption.

The numerical model

The numerical model used to solve the system equations, also has errors associated with it such as parametrisation errors [34]. The model errors in the problems considered for this thesis, can be divided into *errors in the model equations* and *numerical model error*. The former is a result of the model equations for the physical system failing to capture the behaviour of the physical system. One possible reason for this type of error is a lack of understanding of the physical processes in the system. Another is due to the limitations imposed by computational resources. The system of equations that model the physical system may be too complex to solve on the computational resources available, in the time required. Therefore a simplified model may be used instead. This may especially be true in highly non-linear systems of equations.

Numerical model error arises due to errors in the implementation of the numerical methods used to solve the system of equations chosen to model the physical system. It is the effects of this type of error on the results of data assimilation, which are the focus of interest in this thesis. Computers are unable to solve PDEs directly. Therefore approximations such as finite differences are used to approximate PDEs, in

order to obtain a computational solution. These approximations result in errors being introduced into the solution. The accuracy of these models depends upon the resolution of the grid over which the model is implemented.

The function $\mathcal{M}_{l+1,l} : \mathbb{R}^N \to \mathbb{R}^N$, is a possibly non-linear numerical model taking the lth state of the numerical model to its $(l+1)$th state. The vector $\mathbf{x}_l \in \mathbb{R}^N$ contains the state of the numerical model at every grid point in space, for every variable the model solves for, at the time of the lth observation. Strong constraint 4D-Var data assimilation makes the assumption that this model is perfect, despite our discussion in Section 2.1, about the presence of model error. This is not a bad assumption when model error is small in comparison to the other errors affecting 4D-Var. However it can be difficult and expensive to determine the properties of model error [10].

Many of the available data assimilation methods have been developed to account for some of the errors discussed in this Section. However, some methods operate on the assumption that there are no errors in the numerical model for the physical system. The strong constraint variational data assimilation methods operate under this assumption. It is important to understand whether this assumption is realistic and if not, how the methods should be adapted to take account of model errors. The method of 4D-Var data assimilation has become more widely used since the invention of *Incremental 4D-Var data assimilation*, which provides a computationally less expensive implementation of the method. This method is used in applications such as NWP, where the weather is a highly chaotic system. The improved initial condition found through data assimilation, is used to generate a weather forecast. Therefore any errors in this initial condition, can result in a highly erroneous forecast. Hence we choose to investigate the effects of numerical model error on the results of 4D-Var data assimilation. More details on the variables associated with 4D-Var can be found in [2, 11, 16, 18, 20, 21]. In the next Sections we review the method of 4D-Var and its implementation through incremental 4D-Var data assimilation.

2.1.1 4D-Variational data assimilation

The method of 4D-Var data assimilation, implemented in applications such as NWP, solves a particular formulation of the data assimilation problem.

Given a set of observations of a physical system taken over a period of time, a numerical model of the system and a priori information on the initial condition for the physical system, estimate an initial condition for the numerical model that best replicates the true state of the system.

The method of strong constraint 4D-Var data assimilation solves this problem by performing a constrained weighted least-squares minimisation, which calculates an es-

timate of the true initial condition for the physical system. Once this process has been completed, the estimated initial condition can be used in the numerical model, to generate a forecast for the system. The solution to the problem is obtained through the minimisation of the *strong constraint 4D-Var cost function*, $J : \mathbb{R}^N \to \mathbb{R}$, with respect to the initial conditions for the numerical model $\mathbf{x}_0 \in \mathbb{R}^N$,

$$J(\mathbf{x}_0) = (\mathbf{x}_b - \mathbf{x}_0)^T B^{-1}(\mathbf{x}_b - \mathbf{x}_0) + \sum_{l=0}^{L} \left[\mathbf{y}_l - \mathcal{H}_l\left(\mathbf{x}_l\right)\right]^T R_l^{-1} \left[\mathbf{y}_l - \mathcal{H}_l\left(\mathbf{x}_l\right)\right], \quad (2.1)$$

$$\mathbf{x}_{l+1} = \mathcal{M}_{l+1,l}(\mathbf{x}_l). \quad (2.2)$$

The solution $\mathbf{x}_a \in \mathbb{R}^N$ will be termed the *analysis vector*, following the convention of NWP literature, ie: $\nabla J(\mathbf{x}_a) = 0$. It forms a maximum likelihood estimate for the state of the system [2].

The term '*strong constraint*' refers to the fact that the model for the physical system is used as a strong constraint, by assuming that the model is perfect. This is opposed to the '*weak constraint*' formulation which will be described in Section 2.2.1. Strong constraint 4D-Var was first proposed by Le Dimet and Talagrand [35]. This formulation only minimises with respect to the initial condition, rather than for all model states $\mathbf{x}_l$, reducing the number of state variables we need to minimise with respect to [10]. A derivation can be found in papers such as Lorenc [2] and Bouttier and Courtier [18].

The strong constraint 4D-Var cost function

The strong constraint 4D-Var cost function, constructs a constrained weighted least-squares solution to the 4D-Var data assimilation problem. It compares the forecast from the numerical model with observations of the physical system, to create an improved initialisation. The background and observation covariance matrices act as weights in the minimisation process, allowing the process of strong constraint 4D-Var data assimilation to attempt to account for the effects of background, representative and observation errors. When the background estimate is more accurate than the observations, it is given a greater weight in the minimisation. Similarly, when the observations are more accurate than the background estimate, they are given greater weight [36]. Formulating the cost function in this way, making use of the covariance matrices, is an advantage to the method. We need only estimate the covariance matrices for the associated errors, which is generally all we know about them [10].

Johnson et al. [17] show that strong constraint 4D-Var data assimilation can be interpreted as a Tikhonov regularisation problem, by considering the background and observation error covariance matrices to be diagonal matrices such that $B^{-1} = \frac{1}{\sigma_b^2} I_N$ and $R_l^{-1} = \frac{1}{\sigma_o^2} I_m$, for some $m \in \mathbb{N}$ for all l, $\sigma_b, \sigma_0 \in \mathbb{R}^+$. The regularisation parameter $\frac{\sigma_o^2}{\sigma_b^2}$ is investigated to understand the effect of its value on the accuracy of the forecast. This work has shown how finding the balance between the contribution from these er-

rors is important. They performed strong constraint 4D-Var using the two-dimensional Eady model as the physical system, together with perfect observations given additive, uncorrelated observation errors and a background estimate constructed from the true state of the system, with a phase shift. The background errors were assumed uncorrelated also. The results showed that information may be lost if the weighting of the background term is too great. However, if the observations are not weighted enough, the analysis vector is sensitive to observation errors [17].

The method of 3D-Var data assimilation is obtained by performing strong constraint 4D-Var data assimilation, using only one observation at $l = 0$. The advantage of using 3D-Var is that the numerical model does not need to be executed to generate the state of the system over the assimilation window, on each iteration of the minimisation. This makes its implementation less computationally expensive. The analysis vector from this process is then used in the numerical model to generate a forecast for the physical system. This process is repeated for every new set of observations in time. As observations are frequent in time, the forecast doesn't run for very long before a new analysis vector is calculated, creating unrealistic jumps in the state of the system. Strong constraint 4D-Var reduces this problem by using several sets of observations over time, so the unrealistic jumps are much further apart. The forecast presents a smooth prediction for the state of the system [18]. However the downside is the computational cost associated with the process of strong constraint 4D-Var data assimilation.

The high computational costs associated with strong constraint 4D-Var, in part arise due to the need to execute the numerical model on each iteration of the minimisation process. The non-linearity of the physical system adds to this cost by leading to a non-linear numerical model. The observation operators can also be non-linear due to the transformations required to map state variables to observation variables, also adding to computational costs. The computational cost of observation operators is also present in 3D-Var, but the large number of observations associated with 4D-Var increases this further.

If the numerical model and/or the observation operator are non-linear, this results in a cost function which is no longer quadratic. The minimisation process then becomes more computationally expensive. Another disadvantage is that it may find a local minimum if the background estimate does not constrain the solution near the global minimum [10]. The method of *incremental 4D-Var data assimilation* was developed by Courtier et al. [12] to provide a computationally viable implementation of strong constraint 4D-Var for non-linear systems. It also takes advantage of the many minimisation algorithms available for quadratic functions by linearising non-linear operators [10].

2.1.2 Incremental 4D-Variational data assimilation

Incremental 4D-Var data assimilation is an iterative method, allowing the 4D-Var cost function to be minimised efficiently [5]. The process involves an inner and outer loop process. Let $\mathbf{x}_0^{(k)}$ denote the kth estimate for $\mathbf{x}_a$, $k \in \mathbb{N}_0$. Then the outer loop calculates, $\mathbf{x}_0^{(k+1)} = \mathbf{x}_0^{(k)} + \boldsymbol{\delta}\mathbf{x}_0^{(k)}$. Here $\boldsymbol{\delta}\mathbf{x}_0^{(k)} \in \mathbb{R}^N$ is the kth update for the initial condition, identified through the inner loop. The background estimate is usually chosen as $\mathbf{x}_0^{(0)}$. The fully non-linear model is executed with each $\mathbf{x}_0^{(k)}$ and then $\mathbf{d}_l^{(k)} \in \mathbb{R}^N$ is calculated for all $l = 0, \ldots, L$, such that $\mathbf{d}_l^{(k)} = \mathbf{y}_l - \mathcal{H}_l(\mathbf{x}_l^{(k)})$ [5].

The inner loop calculates $\boldsymbol{\delta}\mathbf{x}_0^{(k)}$ through minimising a linearised version of the 4D-Var cost function. The non-linear aspects of the 4D-Var cost function arise due to non-linear numerical models and observation operators. Therefore in order to create an efficient method for minimising the cost function, the lth numerical model and observation operator are linearised about the current state of the numerical model, $\mathbf{x}_l^{(k)}$,

$$\mathcal{H}_l(\mathbf{x}) \approx \mathcal{H}_l(\mathbf{x}_l^{(k)}) + H_l(\mathbf{x} - \mathbf{x}_l^{(k)}), \tag{2.3}$$

$$\mathcal{M}_{l+1,l}(\mathbf{x}) \approx \mathcal{M}_{l+1,l}(\mathbf{x}_l^{(k)}) + M_{l+1,l}(\mathbf{x} - \mathbf{x}_l^{(k)}). \tag{2.4}$$

Here $H_l \in \mathbb{R}^{m_l \times N}$, $M_{l+1,l} \in \mathbb{R}^{N \times N}$, are the Jacobian matrices for the observation operator and numerical model, respectively. This assumption is only valid if the higher-order terms of the expansion can be neglected [18]. The matrices $M_{l+1,l}$ and $M_{l+1,l}^T$ are referred to as the *tangent linear model* (TLM) and the *adjoint model* [10] respectively, in NWP literature, where $\cdot^T$ denotes the matrix transpose. The resultant cost function, minimised to identify $\boldsymbol{\delta}\mathbf{x}_0^{(k)}$ is [5],

$$\begin{aligned}\hat{J}^{(k)}(\boldsymbol{\delta}\mathbf{x}_0^{(k)}) = &\left[\boldsymbol{\delta}\mathbf{x}_0^{(k)} - (\mathbf{x}_b - \mathbf{x}_0^{(k)})\right]^T B^{-1} \left[\boldsymbol{\delta}\mathbf{x}_0^{(k)} - (\mathbf{x}_b - \mathbf{x}_0^{(k)})\right] \\ &+ \sum_{l=0}^{L} \left[H_l \boldsymbol{\delta}\mathbf{x}_l^{(k)} - \mathbf{d}_l^{(k)}\right]^T R_l^{-1} \left[H_l \boldsymbol{\delta}\mathbf{x}_l^{(k)} - \mathbf{d}_l^{(k)}\right].\end{aligned}$$

where the perturbation $\delta x_l^{(k)}$ is calculated using [11],

$$\delta_{l+1}^{(k)} = M_{l+1,l} \delta x_l^{(k)}.$$

This effectively solves the problem of needing to minimise a cost function that is not quadratic. Lawless et al. [37] showed that incremental 4D-Var is equivalent to minimising the original non-linear cost function, using an inexact Gauss-Newton method [11]. Another advantage is that the linearised problem can be solved at a lower spatial resolution and the solution can be returned to the higher resolution problem in the outer loop [5, 11]. A complete algorithm for incremental 4D-Var can be found in the PhD thesis of Haben [5, Section 2.3.1, p. 13]. A discussion on the practical implementation

of 4D-Var can be found in Lawless [11, 18].

When methods such as the pre-conditioned conjugate gradient (PCG) method are used to iteratively minimise the 4D-Var cost function, this requires evaluations of both the cost function and its gradient. The adjoint model is used to efficiently calculate the gradient of the cost function via the discrete adjoint equations [11, 18],

$$\lambda_l = \begin{cases} M_{l+1,l}^T \lambda_{l+1} - H_{l+1,l}^T R^{-1} \left(\mathcal{H}_{l+1,l}(\mathbf{x}_l) - \mathbf{y}_l \right), & \text{for } l = 0, \ldots, L, \\ \mathbf{0}, & \text{for } l = L+1. \end{cases}$$

The vectors $\lambda_l \in \mathbb{R}^N$ are the adjoint variables of the minimisation and can be used to measure the sensitivity of the cost function to changes in inputs [38]. The gradient of the cost function is then calculated by [11, 18],

$$\nabla J(\mathbf{x}_0) = -\lambda_0 + B^{-1} \left(\mathbf{x}_0 - \mathbf{x}_b \right).$$

One advantage of incremental 4D-Var data assimilation is that the cost function of the inner loop is quadratic, so there are a large number of efficient minimisation algorithms available, such as the *conjugate gradient* (CG) method. The minimisation does not require running the fully non-linear models, only requiring knowledge of the tangent-linear and adjoint models [10]. Only a few iterations of the outer loop is required, so the fully non-linear model need only be executed a few times. This results in the outer loop being the most computationally expensive part of incremental 4D-Var [5]. The result of this method should be a "nearly optimal" analysis vector [18]. Lawless at al. [39] showed that the accuracy of results can be dependent on the number of outer loops performed. Information on the implementation of incremental 4D-Var at the Met. Office and ECMWF can be found in [40] and [4] respectively.

2.2 Model error in data assimilation

The derivation of strong constraint 4D-Var data assimilation makes the assumption that the forward model used to solve the model equations is perfect [2]. In order to account for the effects of model error on strong constraint 4D-Var, several different methods have been suggested. One method proposed is to use a modified formulation for the model equations, by augmenting them to account for model errors. This is termed *weak constraint 4D-Var data assimilation*, leading to the original formulation to be termed strong constraint 4D-Var data assimilation. The method of weak-constraint 4D-Var is described in Section 2.2.1.

Le Dimet and Shutyaev [3] performed sensitivity analyses to identify the impact of different forms of error associated with strong constraint 4D-Var, on the accuracy of the optimal solution produced by the analysis vector. They identified that the error in the forecast formed from the analysis vector is most sensitive to observation errors and advocate the use of regularisation to ensure that the prediction error remains stable

[3].

Budd et al. [41] use the Tikhonov regularisation formulation of the strong constraint 4D-Var cost function [17], to apply a mixed total variation $L_1 - L_2$-norm regularisation to the cost function, opposed to the $L_2 - L_2$-norm regularisation the cost function usually possesses ie: the L_2-norm on the background term of the cost function is replaced by an L_1-norm. This approach improved the accuracy of the analysis vector, even in the "presence of sharp fronts and model error" [41]. Zou et al. [42] also make use of penalty functions to reduce the effects of model error in the shallow water equations. They also investigate the effects of incomplete observations on the minimisation process of the cost function and how penalty functions can be used to improve results [42].

Daley [1] aims to calculate the covariance matrix for forecast errors using Kalman filter theory. He sets out clearly that forecast error is comprised of model errors and *predictability errors*. Predictability errors arise due to the errors in the initial condition for the model, in this case, the analysis vector. If the numerical model were perfect, then predictability errors would still arise in the forecast error. The size and scale of these errors depends on the chaotic nature of the physical system. Determining the covariance matrix for the forecast error using Kalman filter theory requires a covariance matrix for model errors. Estimating the model error covariance matrix allows the accuracy of the perfect model assumption in strong constraint 4D-Var data assimilation to be assessed [1]. Ménard and Daley [43] use Kalman smoothing to estimate 4D-Var error statistics.

2.2.1 Weak constraint 4D-Var data assimilation

Weak constraint 4D-Var was first suggested by Sasaki [44] and uses the model equations as a weak constraint for the minimisation process by augmenting the original model equations with a model error correction term. Unfortunately, not much is known about how to formulate model error in the model equations [45]. Vidard et al. [46] consider the following model formulation for weak constraint 4D-Var, introduced by Griffith et al. [45],

$$\begin{aligned} &\tfrac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \mathcal{M}(\mathbf{x}(t)) + \mathcal{T}(\boldsymbol{\eta}(t)), \\ &\tfrac{\mathrm{d}\boldsymbol{\eta}(t)}{\mathrm{d}t} = \Phi(\boldsymbol{\eta}(t), \mathbf{x}(t)) + \boldsymbol{\epsilon}(t), \\ &\mathbf{x}(t_0) = \mathbf{x}_0, \;\; \boldsymbol{\eta}(t_0) = \boldsymbol{\eta}_0, \;\; \boldsymbol{\epsilon}(t_0) = \boldsymbol{\epsilon}_0. \end{aligned} \tag{2.5}$$

Here $\mathbf{x} : [0, \infty) \to \mathbb{R}^N$ is the state of the physical system over time, $N \in \mathbb{N}$; $\mathcal{M} : \mathbb{R}^N \to \mathbb{R}^N$ denotes the model equations for the time evolution of the physical system; $\boldsymbol{\eta} : [0, \infty) \to \mathbb{R}^P$ denotes the model error in $P \leq N$ entries of $\mathcal{M}(\mathbf{x}(t))$, $P \in \mathbb{N}$; $\mathcal{T} : \mathbb{R}^P \to \mathbb{R}^N$ maps the model error in $\boldsymbol{\eta}(t)$ to their corresponding variables in $\mathcal{M}(\mathbf{x}(t))$; $\boldsymbol{\epsilon} : [0, \infty) \to \mathbb{R}^P$ is the unbiased, serially uncorrelated, normally distributed stochastic part of the time evolution of the model error [46]. Estimating model error in this way is consistent with the fact that the model equations are initially posed in continuous form [47]. These equations are then discretised so they can be solved numerically.

The weak-constraint model formulation proposes treating model error in the form of a time evolving function constructed using separate terms to account for deterministic and stochastic model errors. Dee and Da Silva [27] make the point that it is important to account for both deterministic and stochastic errors. Considering only one form of these errors in the model formulation will result in the unaccounted for error affecting the ability of the model to produce the desired improvement [27]. Model errors are assumed to be white noise, uncorrelated in time, however this is generally not the case as the model error at a particular point in time is influenced by the model error at a previous point in time [45].

Derber [48] suggested accounting for the effects of deterministic model errors by choosing, $\boldsymbol{\eta}(t) = \lambda(t)\phi$ ($\boldsymbol{\epsilon}(t) = 0$) so that $\lambda(t)$ is time dependent and ϕ is a spatially dependent variable [47, 48]. The cost function is formulated as for the strong constraint 4D-Var problem, using this $\boldsymbol{\eta}(t)$ in (2.5), neglecting the background term and choosing $R_l^{-1} = I_N$ for all l. The aim is to estimate ϕ using the observations rather than $\mathbf{x}_0$, so that model updates are used to re-calibrate the forecast rather than updates to the initial condition. Trial functions for $\lambda(t)$ were a parabolic function, a constant function and a delta function. The delta function form of $\lambda(t)$ produced a result similar to the strong constraint formulation. An improvement was seen when using either the parabolic or constant functions for $\lambda(t)$ [48].

Zupanski [49] extended the results of Derber [48] using the same weak-constraint model formulation, with the parabolic function for $\lambda(t)$ and a modified cost function. This formulation also resulted in an improvement in the forecast for the considered physical system.

Wergen [50] also uses a constant function for $\boldsymbol{\eta}(t)$, using the same cost function as Derber [48] with an additive correction term for noisy data, whose formulation is developed from prior knowledge. The cost function also has another term added to it minimising the value of the considered non-zero constants in $\boldsymbol{\eta}(t)$. The results revealed that the *root mean square error* (RMSE) of the forecast was improved.

Estimating model error in this way is computationally expensive due to the number of control variables which need to be estimated. This is especially true for NWP [45]. Vidard et al. [51] suggest that the computational cost can be reduced by controlling the direction of $\boldsymbol{\eta}(t)$. This approach was found to improve the results of their weak-constraint 4D-Var data assimilation experiments further [51].

Recent papers have formulated the weak constraint 4D-Var cost function to minimise for both the model error correction term and the initial condition for the numerical model. This allows the analysis vector to be estimated whilst limiting the effects of model error. Suppose the discrete model error is assumed to be purely stochastic such that,

$$\mathbf{x}_{l+1} = \mathcal{M}_{l+1,l}(\mathbf{x}_l) + \boldsymbol{\epsilon}_{l+1}, \tag{2.6}$$

where $\boldsymbol{\epsilon}_l \in \mathbb{R}^N$ is the stochastic model error assumed to be unbiased and a serially

uncorrelated Gaussian random variable with covariance matrix $Q_l \in \mathbb{R}^{N\times N}$ for $l = 0, \ldots, L$ [45, 52]. Then the resulting cost function is minimised to identify $\mathbf{x}_0$ and $\boldsymbol{\epsilon}_l$ for $l = 0, \ldots, L$,

$$\begin{aligned} J(\mathbf{x}_0, \boldsymbol{\epsilon}_0, \ldots, \boldsymbol{\epsilon}_L) &= (\mathbf{x}_b - \mathbf{x}_0)^T B^{-1}(\mathbf{x}_b - \mathbf{x}_0) + \sum_{l=0}^{L} \left[\mathbf{y}_l - \mathcal{H}_l\left(\mathbf{x}_l\right)\right]^T R_l^{-1} \left[\mathbf{y}_l - \mathcal{H}_l\left(\mathbf{x}_l\right)\right] \\ &\quad + \sum_{l=0}^{L} \boldsymbol{\epsilon}_l^T Q_l^{-1} \boldsymbol{\epsilon}_l. \end{aligned} \tag{2.7}$$

This approach is taken by Furbish et al. [52] and Trémolet [53]. The assumptions placed on $\boldsymbol{\epsilon}_l$ are unlikely to be valid in practice due to the time correlated nature of model errors [45]. Gathering information to construct Q_l is infeasible [53]. Approximations for Q_l have been considered, such as αB for some $\alpha \in \mathbb{R}$, but this was found not to be a good approximation [53].

Considering deterministic model errors results in minimising the number of model parameters to be estimated in the cost function. Consider the new discrete model equations,

$$\begin{aligned} \mathbf{x}_{l+1} &= \mathcal{M}_{l+1,l}(\mathbf{x}_l) + \mathcal{T}_l \boldsymbol{\eta}_l, \\ \boldsymbol{\eta}_{l+1} &= \Phi(\mathbf{x}_l, \boldsymbol{\eta}_l), \end{aligned} \tag{2.8}$$

where $\boldsymbol{\eta}_l \in \mathbb{R}^N$ is the deterministic model error and $\mathcal{T}_l \in \mathbb{R}^{N\times P}$ is a matrix mapping the model error correction to model space [45, 46, 47]. As the model error is deterministic, only $\mathbf{x}_0$ and $\boldsymbol{\eta}_0$ need to be estimated to generate a forecast, reducing the number of control variables in the cost function,

$$\begin{aligned} J(\mathbf{x}_0, \boldsymbol{\eta}_0) &= (\mathbf{x}_b - \mathbf{x}_0)^T B^{-1}(\mathbf{x}_b - \mathbf{x}_0) + \sum_{l=0}^{L} \left[\mathbf{y}_l - \mathcal{H}_l\left(\mathbf{x}_l\right)\right]^T R_l^{-1} \left[\mathbf{y}_l - \mathcal{H}_l\left(\mathbf{x}_l\right)\right] \\ &\quad + (\boldsymbol{\eta}_b - \boldsymbol{\eta}_0)^T Q^{-1}(\boldsymbol{\eta}_b - \boldsymbol{\eta}_0). \end{aligned} \tag{2.9}$$

This formulation is considered by Akella and Navon [47], Griffith and Nichols [45] and Vidard et al. [46]. The vector $\boldsymbol{\eta}_b$ is composed of a priori information on $\boldsymbol{\eta}_0$. Gathering prior knowledge on the initial model error is still a problem [46]. Here $Q \in \mathbb{R}^{P\times P}$ is the covariance matrix for the error in $\boldsymbol{\eta}_b$, which is assumed to be a Gaussian random variable [45].

Akella and Navon [47] investigate the impact of choosing $\Phi(\boldsymbol{\eta}(t)) = \beta\boldsymbol{\eta}(t)$ in (2.8), where β is a constant coefficient, on the effects of discretisation errors. As this is a deterministic error, stochastic errors are neglected. Discretisation errors enter into all model variables so $\mathcal{T}(\boldsymbol{\eta}(t)) = \eta(t)$. In order to investigate the effects of choosing a growing, constant or decaying linear function of $\boldsymbol{\eta}(t)$, $\beta > 0$, $\beta = 0$ and $\beta < 0$ are investigated respectively. Accounting for model error improved the results for each choice of $\Phi(\cdot)$, but the constant model error provided the best results [47]. This is surprising as you might expect model error to compound over time so a growing model

error might be appropriate. It is important to note though that different types of error will arise in different physical systems and numerical models, so different model error formulations will be required [45].

2.2.2 Numerical dissipation and dispersion

Numerical model errors in advection problems can lead to physically unrealistic anomalies in the solution that can have wide reaching consequences on the numerical results of the forward model over time [54]. Vukićević et al. [55] examined the numerical values of and errors in the analysis vector generated through strong constraint 4D-Var data assimilation experiments, using three different schemes to solve the 2D linear advection equation. The observations used were assumed to be perfect and three different initial errors were considered for the background estimate in the minimisation of the cost function. The numerical results obtained exhibited behaviours due to the effects of numerical dissipation and dispersion in the advection schemes. The accuracy of the results was also found to be positively correlated to the accuracy of the forward and adjoint models [55]. Both Gerdes et al. [54] and Vukićević et al. [55] discuss the impact of numerically dissipative and dispersive effects from their advection schemes on the results of their 4D-Var experiments. Some of these can be desirable whilst others are not. Hence it is important to understand analytically the impact of numerical dissipation and dispersion on the results of 4D-Var.

Numerical dissipation and *numerical dispersion* will be formally defined in Chapter 3, but occur when wavenumber components of the solution are propagated with an incorrect amplitude and phase speed, respectively. Numerical non-dissipative and dispersive effects on high resolvable wavenumber components, can lead to the introduction of rapidly varying noise in the numerical solution [6]. Schemes have been designed to add artificial numerical dissipation, to reduce these effects. Examples of such schemes can be found in [56, 57]. It is important that a sufficient number of discretisation points be used with these schemes so that noise is removed, but not at the expense of overall accuracy in the numerical solution [6]. However, modifying a scheme to account for one form of error, can result in the magnification of another form of error [58].

2.3 Problem formulation

Here we wish to consider the effects of numerical model error on strong constraint 4D-Var data assimilation. In order to identify these effects, we remove all other forms of error to perform our analysis. This involves the use of perfect observations. Once this has been completed, other forms of error can then be re-introduced, to investigate how they behave together. Since the forecast from the analysis vector obtained through strong constraint 4D-Var, has been shown to be most sensitive to observation errors [3], we will re-introduce them to the problem after our initial analysis.

We consider the 1D and 2D linear advection equations, as well as the 2D linearised shallow water equations, together with circulant boundary conditions and initial conditions, as our physical systems of interest. The 1D linear advection problem provides a physical system which looks deceptively simple. Our reasons for choosing to investigate this physical system were set out in Chapter 1. The 2D linear advection problem aims to expand the results from the 1D linear advection equation. The 2D linearised shallow water problem then increases the complexity further by considering a 2D coupled system of PDEs.

Examining linear problems is relevant for adjoint methods and the tangent linear model in incremental 4D-Var. Numerical models can be constructed to solve the considered physical systems, by using finite differences to approximate derivatives [14]. This forms finite difference schemes for solving the model equations, which introduce deterministic numerical model errors into the solution of the minimisation of the cost function. There are many schemes available to solve each of these systems, so it is possible to compare the different effects introduced by different properties of the schemes.

In order to fully investigate the effects of this form of deterministic numerical model error, all other errors present in the problem initially need to be removed. Therefore, the background term of the cost function is neglected as in Griffith and Nichols [45] and Vukićević et al. [55] in order to allow the full impact of deterministic numerical model error to be seen. Taking an observation at time $l = 0$ acts to regularise the problem so that it remains well-posed. This is demonstrated in Section 3.10. Therefore we are able to remove the effects of background errors from the problem.

Define $\tilde{\mathbf{y}}_l \in \mathbb{R}^{m_l}$ as the lth perfect observation of the physical system (that is, no observation errors). Also define $\boldsymbol{\epsilon}_l \in \mathbb{R}^{m_l}$ as the observation error in the lth observation, such that the lth observation of the true physical system is given by $\mathbf{y}_l = \tilde{\mathbf{y}}_l + \boldsymbol{\epsilon}_l$. We assume $\boldsymbol{\epsilon}_l$ to be an independently and identically distributed (iid) Gaussian random variable such that $\boldsymbol{\epsilon}_l \sim \mathcal{N}(\mathbf{0}, \sigma_o^2 I_N)$, leading to $R_l = \sigma_o^2 I_N$ for all l, $\sigma_o \in \mathbb{R}^+$. We take observations at every spatial and temporal grid point of the numerical model, of all state variables of the numerical model. This eliminates any need to interpolate and transform model states using the observation operator, removing representative errors. The model equations are assumed to be the equations of the physical system, so there are no model errors due to inaccurate model equations. Hence $m_l = N$ and $\mathcal{H}_l = I_N$ the $N \times N$ identity matrix, for all $l = 0, \ldots, L$. These assumptions result in the following cost function,

$$J(\mathbf{x}_0) = \frac{1}{\sigma_o^2} \sum_{l=0}^{L} \left[\mathbf{y}_l - \mathcal{M}_{l,0}(\mathbf{x}_0)\right]^T \left[\mathbf{y}_l - \mathcal{M}_{l,0}(\mathbf{x}_0)\right]. \tag{2.10}$$

We remind the reader that $\mathcal{M}_{l,0}$ is the model operator mapping the state of the physical system from time t_0 to time t_l.

Initially the problem will be investigated in the absence of observation errors. If R_l^{-1} was chosen so as to reflect the statistical properties of numerical model error,

then at this initial point we know virtually nothing about these statistics, so we choose $R_l^{-1} = I_N$ for all l by taking $\sigma_o = 1$. This choice gives each set of observations an equal weighting, assuming nothing about the error statistics of the numerical model. These model variables were also chosen by Daley [20], Griffith and Nichols [45] and Vukićević et al. [55]. Later observation errors will be reintroduced to the problem.

We also define the following sets in order to remove any ambiguity in the results of this thesis,

$$\mathbb{Q}^+ = \{x \in \mathbb{Q} \,|\, x > 0\}, \tag{2.11}$$

$$\mathbb{R}^+ = \{x \in \mathbb{R} \,|\, x > 0\}. \tag{2.12}$$

CHAPTER 3

The 1D Linear Advection Problem

In this chapter, the strong constraint 4D-Var data assimilation problem set out in Section 2.3 is considered in the absence of observation errors. Therefore only the effects of numerical model error on the analysis vector will be investigated. The physical system of interest in this Chapter is the *1D linear advection equation* together with circulant boundary conditions and initial condition. The d-dimensional linear advection equation is derived from the continuity equation, as demonstrated by Rood [58], $d \in \mathbb{N}$. It can be used to model the movement of pollutants in rivers and provides the opportunity to study the advection of particles without diffusion [13]. Despite pollutant motion being three-dimensional in reality, it may be sufficient to consider the one- or two-dimensional linear advection equation.

We choose the 1D linear advection equation together with circulant boundary conditions as our physical system, for the reasons discussed in Chapter 1. The use of circulant boundary conditions simplifies our analysis by allowing the use of Fourier series methods. There are many finite difference schemes which can be derived for solving this system, introducing numerical model error through the approximation of derivatives by finite differences. As this system is defined by a single linear PDE, the numerical model error introduced can easily be classified as a form of *numerical dissipation* and/or *numerical dispersion*. The aim of this Chapter is to determine the impact of these forms of error on the contribution of resolvable and unresolvable wavenumber components, to the analysis vector in comparison to the discrete sample of the true initial condition we wish to recover.

The chapter begins by reviewing *Fourier series* and the *discrete Fourier transform (DFT)* for one-dimensional systems as they will be useful tools for our analysis. The Upwind, Preissman Box and Lax-Wendroff schemes will be used as the forward models of interest. Through the construction of a Fourier series for the numerical solutions of these schemes, numerical dissipation and numerical dispersion are defined along with the concept of *aliasing errors*, allowing the numerical model error introduced by the

schemes to be classified.

The considered strong constraint 4D-Var problem requires the use of perfect observations. We discuss the options available for generating these observations numerically and algebraically. To this end, we develop the *Modified Numerical Implementation of the Method of Characteristics* (MNIMC) scheme, a numerically non-dissipative and non-dispersive scheme with respect to the resolvable wavenumber components of the numerical solution. This scheme is used to define perfect observations algebraically and metrics for determining the dissipative and dispersive properties of finite difference schemes for solving our physical system. These properties are found to be dependent on the value of the CFL number $h \in \mathbb{R}^+$ for the physical system.

The Chapter continues by taking a spectral approach to understanding the effects of numerical model error on the analysis vector, in the absence of observation errors. We examine the analytical and visual properties of the analysis vector, when the true initial condition is a step function, demonstrating a shock profile. This initial condition is constructed from high wavenumber components, allowing us to conduct a similar analysis to the "spike test" discussed by Durran [6]. We investigate how the numerically dissipative and/or dispersive properties of the schemes, as well as the number of sets of observations in the assimilation window, affects the results. The Chapter concludes with a discussion of the results.

3.1 The physical system

Consider the 1D linear advection equation, for the function $u : \mathbb{R} \times [0, \infty) \to \mathbb{R}$, $(x, t) \mapsto u(x, t)$, together with circulant boundary conditions and initial condition, $u_0 : [0, 1) \to \mathbb{R}$, $x \mapsto u_0(x)$,

$$\begin{aligned} &u_t(x,t) + \mu u_x(x,t) = 0, && x \in [0,1), \;\; t > 0, \\ &u(x,t) = u(x+1,t), && x \in \mathbb{R}, \;\; t \geq 0, \\ &u(x,0) = u_0(x), && x \in [0,1). \end{aligned} \tag{3.1}$$

This is a linear, hyperbolic, one-dimensional PDE problem [6]. Here the *wave speed* $\mu \in \mathbb{R}$ remains constant. The 1D linear advection equation is also considered in the context of data assimilation by Freitag et al. [41], Griffith and Nichols [45] and Haben [5]. It is important to note that the scalar x is the spatial dimension whilst the vectors $\{\mathbf{x}_l\}_{l=0}^{L}$ defined in Section 2.3 denote the state of the numerical model.

The solution to this problem is, $u(x, t) = u(x - \mu t, 0) = u_0([x - \mu t]_1)$ [59]. This preserves the shape of the initial condition over time and propagates it through space with speed μ. Here $[\cdot]_1$ denotes modulo one.

Problem (3.1) can be solved numerically using a finite difference scheme as the forward model. These find a numerical approximation to the analytic solution, introducing numerical model error, by using finite differences to approximate derivatives.

This error will be considered in the form of *numerical dissipation* and *numerical dispersion* [6]. In order to introduce the concept of numerical dissipation and dispersion and the *discrete Fourier transform* [60], it is helpful to consider the functions $u(x,t)$ and $u_0(x)$ in terms of their *Fourier series*. To this end, the next section reviews Fourier series and their properties using a general function.

3.2 1D Fourier series

Initially consider a general function $f : \mathbb{R} \times [0,\infty) \to \mathbb{R}$, $(x,t) \mapsto f(x,t)$, such that $f(x+T,t) = f(x,t)$ for all t, for finite $T \in \mathbb{R}^+$, ie: f is T-periodic in space. The exponential form of the Fourier series for $f(x,t)$ is given by $S : \mathbb{R} \times [0,\infty) \to \mathbb{R}$, such that $(x,t) \mapsto S(x,t)$ [60, 61, 62],

$$f(x,t) \sim S(x,t) = \sum_{k=-\infty}^{\infty} a_k(t) e^{\frac{2\pi i k x}{T}}, \tag{3.2}$$

where

$$a_k(t) = \frac{1}{T} \int_0^T f(x,t) e^{\frac{-2\pi i k x}{T}} \mathrm{d}x. \tag{3.3}$$

The meaning of the $\sim$ operator will be discussed in the following section, as it relates to the convergence of Fourier series. Equations (3.2) and (3.3) construct an infinite sum representation of $f(x,t)$, from the superposition of Fourier basis functions given by $e^{\frac{2\pi i k x}{T}}$. These functions are orthonormal basis functions [60] and can be viewed in terms of sine and cosine functions,

$$e^{\frac{2\pi i k x}{T}} = \cos\left(\frac{2\pi k x}{T}\right) + i \sin\left(\frac{2\pi k x}{T}\right).$$

This Fourier basis function has a sinusoidal form with wavelength (period in space) $\frac{T}{k}$, which leads to a wavenumber (frequency in space) of $\frac{k}{T}$ and an angular wavenumber of $\frac{2\pi k}{T}$. By dividing k by T, the Fourier basis functions all complete an integer number of wavelengths in spatial length T.

The coefficients of the Fourier series in (3.2) are time dependent. This is because, given $f(x,t)$ for any fixed time t, a Fourier series in space can be constructed for this function using the same Fourier basis functions. However the difference between each Fourier series is the contribution of each Fourier basis function in the construction of $f(x,t)$, for fixed time t. This is determined by the coefficient of the basis function. In order to move through time between each of these Fourier series, it is the coefficients of the Fourier series which need to evolve with time. Evaluating $S(x,t)$ at some (x,t) sums the Fourier basis functions evaluated at x, weighted by their relative coefficient at time t. We term the kth Fourier basis function, multiplied by its corresponding coefficient in the considered Fourier series, the kth wavenumber component of the Fourier series $k \in \mathbb{Z}$.

The coefficients $a_k(t)$ are determined by (3.3), using the orthonormal properties of the Fourier basis functions. As $f(x,t)$ is a real-valued function and $e^{\frac{2\pi ikx}{T}}$ and $e^{\frac{-2\pi ikx}{T}}$ are complex conjugates, the coefficients have the property that $\overline{a_k(t)} = a_{-k}(t)$ for all $k \in \mathbb{Z}$ and t. As a result, for all $k \in \mathbb{Z} \setminus \{0\}$, the kth wavenumber component and the $(-k)$th wavenumber component can be summed to create a real valued function [60],

$$a_k(t)e^{\frac{2\pi ikx}{T}} + a_{-k}(t)e^{\frac{2\pi i(-k)x}{T}} = 2Re[a_k(t)]\cos\left(\frac{2\pi kx}{T}\right) - 2Im[a_k(t)]\sin\left(\frac{2\pi kx}{T}\right). \quad (3.4)$$

The wavenumber component for $k = 0$ is also a real function. It is these quantities that we will refer to as the *real wavenumber components* of the Fourier series. The real wavenumber components have wavenumbers $k \in \mathbb{N}_0$ and ensure that the Fourier series of a real function is real. We consider these real wavenumber components in Sections 3.10.1-3.10.5.

The function $u(x,t)$ in problem (3.1) is 1-periodic, so $T = 1$. Then define the Fourier series of $u(x,t)$ by the following,

$$u(x,t) \sim \sum_{k=-\infty}^{\infty} b_k(t)e^{2\pi ikx}, \text{ where } b_k(t) = \int_0^1 u(x,t)e^{-2\pi ikx}\mathrm{d}x. \quad (3.5)$$

As the function $u_0(x)$ is only defined on $[0,1)$, the Fourier series for $u_0(x)$ forms a Fourier series for the 1-periodic extension of $u_0(x)$,

$$u_0(x) \sim \sum_{k=-\infty}^{\infty} c_k e^{2\pi ikx}, \text{ where } c_k = \int_0^1 u_0(x)e^{-2\pi ikx}\mathrm{d}x. \quad (3.6)$$

As the function $u(x,0)$ is defined as the 1-periodic extension of $u_0(x)$, they are represented by the same Fourier series. Therefore $b_k(0) := c_k$ for all $k \in \mathbb{Z}$.

Whilst defining the Fourier series in this Section, we have described them as a representation for a function, rather than as equal to the function. In order for a series representation of a function to have the potential to be equal to that function, we require the series to be convergent to the function. The following section describes the conditions under which a Fourier series is convergent.

3.2.1 Convergence of Fourier series

The Fourier series in (3.2) is a representation of the function $f(x,t)$. Consider the *truncated Fourier series* for $f(x,t)$, $S_J : \mathbb{R} \times [0,\infty) \to \mathbb{R}$ such that [60],

$$S_J(x,t) = \sum_{k=-J}^{J} b_k(t)e^{\frac{2\pi ikx}{T}}, \quad (3.7)$$

for some $J \in \mathbb{N}_0$. As $J \to \infty$, $S_J(x,t) \to S(x,t)$. The truncated Fourier series can be used to prove that under certain conditions $S_J(x,t) \to f(x,t)$ as $J \to \infty$, hence

$S(x,t) = f(x,t)$. These conditions are set out in the following Theorem, adapted from Churchill and Brown [63], so the function $f(x,t)$ is considered at some fixed time t.

Theorem 3.1. *[63] Consider the T-periodic function $f(x,t)$ in Section 3.2 for fixed time t. Suppose the function is piecewise continuous over $(0,T)$. If the left- and right-hand derivatives of $f(x,t)$ exist for some $x_0 \in \mathbb{R}$, then the Fourier series for $f(x,t)$ in Equation (3.2) is such that,*

$$S(x_0,t) = \frac{1}{2}\left[\lim_{x\to x_0^+} f(x,t) + \lim_{x\to x_0^-} f(x,t)\right]. \tag{3.8}$$

A proof for this Theorem can be found in Churchill and Brown [63]. Under the conditions of this Theorem, the Fourier series of $f(x,t)$ converges uniformly to the function $f(x,t)$, for some fixed time t [63]. The statement of this Theorem is given in various forms throughout the literature, see [60, 61, 62, 63]. The conditions on $f(x,t)$ in Theorem 3.1 are sometimes referred to as *Dirichlet's conditions* [61, 62].

Theorem 3.1 provides sufficient conditions for the convergence of Fourier series [62]. In other words, not every function that has a convergent Fourier series, satisfies this Theorem. There are currently no necessary and sufficient conditions for the convergence of Fourier series [62]. Suppose $f(x,t)$ satisfies Theorem 3.1, then if $f(x,t)$ is continuous at (x_0,t) for fixed t,

$$S(x_0,t) = f(x_0,t),$$

and if $f(x,t)$ is discontinuous at (x_0,t) for fixed t,

$$S(x_0,t) = \frac{1}{2}\left[\lim_{x\to x_0^+} f(x,t) + \lim_{x\to x_0^-} f(x,t)\right].$$

This property is denoted by the $\sim$ operator [60],

$$f(x,t) \sim \sum_{k=-\infty}^{\infty} a_k(t) e^{\frac{2\pi i p x}{T}}. \tag{3.9}$$

If $f(x,t)$ satisfies the conditions for convergence in Theorem 3.1 for all (x,t), then the Fourier series is described as *convergent*. In the next section, we examine the types of discontinuity that can be present within a function that satisfies the convergence properties set out in Theorem 3.1.

3.2.2 Fourier series for discontinuous functions

Consider the function $f(x,t)$ defined in Section 3.2. Suppose the function $f(x,t)$ possesses a discontinuity in $[0,T)$ at some fixed time t^*. Then this discontinuity appears throughout $f(x,t^*)$ due to the T-periodicity of the function in space. We can classify the discontinuity into one of the following types:

- An *infinite discontinuity*; there exists $x_0 \in [0, T)$ such that $\lim_{x \to x_0^+} f(x, t^*)$ and/or $\lim_{x \to x_0^-} f(x, t^*)$ is equal to $\pm\infty$ [61].
- An *undefined discontinuity*; there exists $x_0 \in [0, T)$ such that $\lim_{x \to x_0^+} f(x, t^*)$ and/or $\lim_{x \to x_0^-} f(x, t^*)$ is finite but undefined. For example, if the function is bounded but is highly oscillatory on one side of the limit and a constant on the other side, then the limit is undefined on the oscillatory side [61].
- A *jump discontinuity*; there exists $x_0 \in [0, T)$ such that $\lim_{x \to x_0^+} f(x, t^*)$, $\lim_{x \to x_0^-} f(x, t^*)$ and $f(x_0, t^*)$ all exist and are all finite, but not necessarily equal [61]. This includes point discontinuities.

If we consider each type of discontinuity under the conditions of Theorem 3.1, we find that any function containing an infinite or undefined discontinuity, does not satisfy the conditions of the Theorem. The condition on the existence of the left- and right-hand derivatives ensures that undefined discontinuities cannot be present. The piecewise continuous condition, together with the requirement for the existence of the left- and right-hand derivatives, results in the function being bounded and prohibiting infinite discontinuities. Functions containing jump discontinuities can satisfy the conditions of Theorem 3.1.

Consider the truncated Fourier series in Equation (3.7), for the function $f(x, t)$. Suppose $f(x, t)$ possesses a jump discontinuity at $x_0 \in [0, T)$, at some fixed time t^*. Then if $f(x, t^*)$ satisfies the conditions of Theorem 3.1, $f(x, t^*)$ has a convergent Fourier series. This means that as $J \to \infty$, $S_J(x, t^*) \to f(x, t^*)$ for x where $f(x, t^*)$ is continuous and $S_J(x_1, t^*) \to \frac{1}{2}\left[\lim_{x \to x_1^-} f(x, t^*) + \lim_{x \to x_1^+} f(x, t^*)\right]$ where $x_1 = x_0 \pm sT$ is a point of discontinuity, $s \in \mathbb{Z}$. As the Fourier series for $f(x, t)$ is a continuous function, unlike the function $f(x, t)$, it is not possible for the Fourier series to form the discontinuities present in the function. We have seen this through the convergence of the Fourier series to the midpoint of any jump discontinuity. This means that as $J \to \infty$, oscillations in the Fourier series can be seen about the points of discontinuity. These oscillations are known as *Gibb's phenomenon* and diminish as J increases [60].

Now we are aware of the properties of a Fourier series, we can use them to make sense of the discrete version, the *Discrete Fourier series*. The discrete Fourier series is an important tool for analysing discrete problems. Just as we can represent a continuous function using a Fourier series, a discrete Fourier series can be used to represent the state of a discrete system, such as that produced by a numerical model solving problem (3.1). This representation allows us to take a spectral approach to analysing the effects of numerical model error on the accuracy of the results from our strong constraint 4D-Var problem.

3.3 Finite difference scheme formulation

There are many finite difference schemes available to numerically solve problem (3.1). Rather than studying them all, we consider the Upwind, Preissman Box and Lax-Wendroff schemes, three finite difference schemes that can be used to solve system (3.1). These are 'representative schemes' chosen due to their numerically dissipative and dispersive properties. Others schemes can be found in Le Veque [59].

In order to define a finite difference scheme over our domain $[0, 1]$, we require the following assumptions.

Assumption 3.2. *Divide the domain* $[0, 1]$ *into* $N_x + 1$ *equally spaced mesh points,* $N_x \in \mathbb{N}$. *This gives a grid spacing of* $\Delta x = \frac{1}{N_x}$ *and grid points* $x_j = j\Delta x$, $j = 0, \ldots, N_x$. *Define the time step* $\Delta t \in \mathbb{R}^+$ *for the finite difference scheme and* $t^n = n\Delta t$ *for* $n \in \mathbb{N}_0$. *Let* U_j^n *be the numerical solution at* (x_j, t^n), *such that* $U_j^n \approx u(x_j, t^n)$, *for* $j = 0, \ldots, N_x$ *and* $n \in \mathbb{N}$. *When* $n = 0$, U_j^0 *is created by sampling* $u(x, 0)$, *such that* $U_j^0 := u(x_j, 0)$, *for* $j = 0, \ldots, N_x - 1$. *Define the vector* $\mathbf{U}^n \in \mathbb{R}^{N_x}$ *where the* j*th element of* $\mathbf{U}^n$ *is defined by* $\{\mathbf{U}^n\}_j := U_{j-1}^n$, *for* $j = 1, \ldots, N_x$. *Also, define* $h := \frac{|\mu|\Delta t}{\Delta x}$, *the* CFL number *[14].*

Then the considered schemes are defined by the following schematic equations when $\mu > 0$:

- The Upwind scheme (explicit scheme) [59],

$$U_j^{n+1} = hU_{j-1}^n + (1 - h)U_j^n. \tag{3.10}$$

- The Preissman Box scheme (implicit scheme) [64],

$$(1 - h)U_{j-1}^{n+1} + (1 + h)U_j^{n+1} = (1 + h)U_{j-1}^n + (1 - h)U_j^n. \tag{3.11}$$

- The Lax-Wendroff scheme (explicit scheme) [59],

$$U_j^{n+1} = \frac{h}{2}(h + 1)U_{j-1}^n + (1 - h^2)U_j^n + \frac{h}{2}(h - 1)U_{j+1}^n. \tag{3.12}$$

As a result, we restrict our analysis to the case $\mu > 0$. These finite difference schemes can be applied to the discrete sample of the state of the system found in the vector $\mathbf{U}^n$. This is achieved by constructing a matrix $M \in \mathbb{R}^{N_x \times N_x}$ using the schematic equations, that can be used to post-multiply $\mathbf{U}^n$ to create $\mathbf{U}^{n+1}$, ie: $\mathbf{U}^{n+1} = M\mathbf{U}^n$. This advances the numerical solution at each grid point in space, forward Δt in time and results in $N = N_x$, where N was defined in Section 3.3. Due to the circulant boundary conditions of problem (3.1), the matrix M implementing any of these schemes is circulant [65] and $u(x_{N_x}, t^n) = u(x_0, t^n)$ for all n, hence $U_{N_x}^n = U_0^n$ for all $n \in \mathbb{N}_0$. In the case of the

Upwind scheme,

$$M = \begin{bmatrix} 1-h & 0 & & \cdots & h \\ h & 1-h & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & h & 1-h & 0 \\ 0 & \cdots & 0 & h & 1-h \end{bmatrix}.$$

We now study the eigenvalues and eigenvectors of the matrices implementing the Upwind, Preissman Box and Lax-Wendroff schemes, through the *1D discrete Fourier transform* (DFT).

3.3.1 The 1D discrete Fourier transform

The vector $\mathbf{U}^n$ is N_x-dimensional, so can be constructed from the N_x vectors of the 1D DFT basis [60] $\{\mathbf{v}_p\}_{p=1}^{N_x}$, such that,

$$\{\mathbf{v}_p\}_j := \frac{1}{\sqrt{N_x}} e^{\frac{2\pi i(p-1)(j-1)}{N_x}} = \frac{1}{\sqrt{N_x}} e^{2\pi i(p-1)x_{j-1}}, \quad p, j = 1, \ldots, N_x, \tag{3.13}$$

is the jth element of the pth vector, $\mathbf{v}_p \in \mathbb{C}^{N_x}$. This is the $(p-1)$th Fourier basis function, sampled at x_{j-1}, with amplitude $\frac{1}{\sqrt{N_x}}$. The vectors form an orthonormal basis for $\mathbb{R}^{N_x}$ [60], ie: $\mathbf{v}_p^* \mathbf{v}_q = \delta_{p,q}$. Here $\cdot^*$ denotes Hermitian, ie: $\overline{\cdot}^T$. The numerical solution is then constructed from N_x Fourier basis functions. These vectors form an orthonormal set of eigenvectors for the matrix M, for all three schemes. As a result, an eigenvalue decomposition of the matrix M for each scheme can be constructed using these eigenvectors,

$$M = V \Lambda V^{-1} = V \Lambda V^*. \tag{3.14}$$

The matrix $V \in \mathbb{C}^{N_x \times N_x}$ is constructed from the 1D DFT eigenbasis such that the pth column of V is $\mathbf{v}_p$. It is a unitary matrix, ie: $V^* V = V V^* = I_{N_x}$, the $N_x \times N_x$ identity matrix. The matrix $\Lambda \in \mathbb{C}^{N_x \times N_x}$ is the diagonal matrix of eigenvalues corresponding to the eigenvectors in the matrix V, for the chosen scheme. The eigenvalue $\lambda_p \in \mathbb{C}$ corresponding to $\mathbf{v}_p$ is found in $\{\Lambda\}_{p,p}$ for $p = 1, \ldots, N_x$. The eigenvalues for each scheme are scheme dependent whilst the eigenvectors are scheme independent. Using (3.14) and the unitary property of the matrix V,

$$\mathbf{U}^n = M^n \mathbf{U}^0 = V \Lambda^n V^* \mathbf{U}^0. \tag{3.15}$$

Constructing the discrete sample of the initial condition $\mathbf{U}^0$, from the eigenvectors $\{\mathbf{v}_p\}_{p=1}^{N_x}$ results in,

$$\mathbf{U}^0 = \sum_{p=1}^{N_x} \alpha_p \mathbf{v}_p, \tag{3.16}$$

for some $\alpha_p \in \mathbb{C}$. Equation (3.16) is a *discrete Fourier series* for $\mathbf{U}^0$. This can be seen

by examining an element of $\mathbf{U}^0$ using (3.16). Applying $\mathbf{v}_q^*$ to (3.16) results in,

$$\mathbf{v}_q^*\mathbf{U}^0 = \sum_{p=1}^{N_x} \alpha_p \mathbf{v}_q^*\mathbf{v}_p = \sum_{p=1}^{N_x} \alpha_p \delta_{p,q} = \alpha_q, \tag{3.17}$$

for $q = 1, \ldots, N_x$, by the orthonormality of the eigenvectors. This identifies the coefficients for the eigenvectors of the 1D DFT basis in the construction of $\mathbf{U}^0$, by applying the 1D DFT to $\mathbf{U}^0$. The *1D Inverse DFT* (IDFT) applies the reverse of this. Given the coefficients of the 1D DFT basis, the 1D IDFT constructs the state of the system at each regularly spaced grid point in space. That is, given $\{\alpha_p\}_{p=1}^{N_x}$, the 1D IDFT constructs $\mathbf{U}^0$. This is implemented by applying $\mathbf{v}_j^T$ to the vector of coefficients, creating a discrete Fourier series to represent the state of the system at x_{j-1},

$$\mathbf{v}_j^T[\alpha_1, \ldots, \alpha_{N_x}]^T = \frac{1}{\sqrt{N_x}} \sum_{p=1}^{N_x} \alpha_p \{\mathbf{v}_j\}_p = \frac{1}{\sqrt{N_x}} \sum_{p=1}^{N_x} \alpha_p e^{\frac{2\pi i(p-1)(j-1)}{N_x}} = \{\mathbf{U}^0\}_j, \tag{3.18}$$

for $j = 1, \ldots, N_x$.

As the vectors $\{\mathbf{v}_p\}_{p=1}^{N_x}$ make up the columns of the matrix V, applying V^* to $\mathbf{U}^0$ results in the vector of coefficients for the discrete Fourier series of $\mathbf{U}^0$, $[\alpha_1, \alpha_2, \ldots, \alpha_{N_x}]^T$. This is the matrix form of the 1D DFT, which identifies the coefficients of the discrete Fourier series for any vector $\mathbf{z} \in \mathbb{R}^{N_x}$. When the vector of coefficients is obtained, the matrix V can be applied to recover $\mathbf{z}$, ie: $VV^*\mathbf{z} = \mathbf{z}$. This gives that as V^* is the matrix representation of the 1D DFT, the matrix V is the matrix representation of the 1D IDFT [60].

Define the operator $\mathcal{F} : \mathbb{R}^{N_x} \to \mathbb{C}^{N_x}$, $\mathbf{z} \mapsto \mathcal{F}(\mathbf{z}) = V^*\mathbf{z}$, to implement the matrix form of the 1D DFT. Also denote the pth element of $\mathcal{F}(\mathbf{z})$, the coefficient for the vector $\mathbf{v}_p$ in the construction of $\mathbf{z}$, by $\mathcal{F}_p(\mathbf{z}) = \{V^*\mathbf{z}\}_p = \mathbf{v}_p^*\mathbf{z}$. Using this definition in (3.17) results in $\mathcal{F}_p(\mathbf{U}^0) = \alpha_p$. A comprehensive discussion of the DFT, including its various interpretations, can be found in Briggs and Henson [60].

Consider two one-periodic functions $f, g : \mathbb{R} \to \mathbb{R}$ such that $x \mapsto f(x)$ and $x \mapsto g(x)$ respectively. Suppose these functions are defined so that they are equal at the equally spaced mesh points of our finite different schemes, ie: $f(x_j) = g(x_j)$ for all $j = 0, \ldots, N_x - 1$, and not equal at the remaining points over $[0, 1)$. Then we can see that as the discrete Fourier series for each function is constructed using only the sample points of these functions, the functions have the same discrete Fourier series. This means that the function corresponding to a discrete Fourier series, is non-unique [66]. This property can be useful as we will see in Lemmas 4.3 and 5.10.

Consider the nth state of the numerical model, given by $\mathbf{U}^n = M^n\mathbf{U}^0$. Applying

the 1D DFT to $\mathbf{U}^n$ results in,

$$\begin{aligned} & \mathcal{F}(\mathbf{U}^n) = V^* V \Lambda^n V^* \mathbf{U}^0 = \Lambda^n \mathcal{F}(\mathbf{U}^0), \\ \Rightarrow \quad & \mathcal{F}_p(\mathbf{U}^n) = \lambda_p^n \mathcal{F}_p(\mathbf{U}^0) = \lambda_p^n \alpha_p, \end{aligned} \tag{3.19}$$

for $p = 1, \ldots, N_x$. This gives that the coefficient for the eigenvector $\mathbf{v}_p$, in the construction of $\mathbf{U}^n$ is $\lambda_p^n \alpha_p$ for $p = 1, \ldots, N_x$. This shows that it is the eigenvalues of the matrix M which propagate the state of the system forward Δt through time. As a result, any errors introduced in the propagation of the system, are introduced by errors in these eigenvalues.

Examining the eigenvectors of the 1D DFT basis, it can be seen that the eigenvectors $\mathbf{v}_p$ and $\mathbf{v}_{N_x-p+2}$ are complex conjugates for $p = 2, \ldots, N_x$. It should be noted that when N_x is even, $\mathbf{v}_{\frac{N_x}{2}+1}$ is real. The eigenvector $\mathbf{v}_1$ is always real. This means that when the 1D DFT basis is used to construct the state of a real system, as is true for $\mathbf{U}^n$, the coefficients for these eigenvectors are complex conjugates. This can be seen in (3.17), where α_q and α_{N_x-q+2} are complex conjugates as $\mathbf{v}_q$ and $\mathbf{v}_{N_x-q+2}$ are complex conjugates for $q = 2, \ldots, N_x$. As a consequence, $\overline{\mathcal{F}_p(\mathbf{z})} = \mathcal{F}_{N_x-p+2}(\mathbf{z})$ for $p = 2, \ldots, N_x$, for some vector $\mathbf{z} \in \mathbb{R}^{N_x}$. As a consequence, by examining (3.19), we find that $\lambda_1 \in \mathbb{R}$ and $\overline{\lambda}_p = \lambda_{N_x-p+2}$ for $p = 2, \ldots, N_x$. Hence we consider the eigenvalues in polar co-ordinate form $\lambda_p = |\lambda_p| e^{i\theta_p}$, $\theta_p \in [-\pi, \pi)$ where $\theta_1 = 0$ and $-\theta_p = \theta_{N_x-p+2}$ for $p = 2, \ldots, N_x$.

Summing the complex conjugate wavenumber components of a discrete Fourier series results in a real wavenumber component,

$$\begin{aligned} & \alpha_p [\mathbf{v}_p]_j + \alpha_{N_x-p+2} [\mathbf{v}_{N_x-p+2}]_j \\ = \quad & \frac{\alpha_p}{\sqrt{N_x}} e^{\frac{2\pi i (p-1)(j-1)}{N_x}} + \frac{\overline{\alpha_p}}{\sqrt{N_x}} e^{\frac{2\pi i (N_x-p+1)(j-1)}{N_x}} \\ = \quad & \frac{1}{\sqrt{N_x}} \left\{ 2Re[\alpha_p] \cos\left(2\pi(p-1)x_{j-1}\right) - 2Im[\alpha_p] \sin\left(2\pi(p-1)x_{j-1}\right) \right\}, \end{aligned} \tag{3.20}$$

for $j = 1, \ldots, N_x$. The result is that the state of the system is constructed from $\left\lfloor \frac{N_x}{2} \right\rfloor + 1$ real wavenumber components. Here $\lfloor \cdot \rfloor$ denotes the *floor function*. In particular this is the $(p-1)$th real wavenumber component of a Fourier series, sampled at grid point x_{j-1}. This can be seen by comparing Equation (3.20) with Equation (3.4). The link between the coefficients of the Fourier series and the coefficients of the discrete Fourier series, will be made in Section 3.4.1.

We have previously discussed that the eigenvector $\mathbf{v}_p$ corresponds to the $(p-1)$th Fourier basis function. However in (3.20), the wavenumber components corresponding to $p = \left\lfloor \frac{N_x}{2} \right\rfloor + 2, \ldots, N_x$ have played the role of the negative wavenumber components of the Fourier series, as they form the conjugate pairs of the wavenumber components

for $p = 2, \ldots, \lfloor \frac{N_x}{2} \rfloor + 1$. This can be explained by noticing that,

$$[\mathbf{v}_{N_x-p+2}]_j = \frac{1}{\sqrt{N_x}} e^{\frac{2\pi i(N_x-p+1)(j-1)}{N_x}} = \frac{1}{\sqrt{N_x}} e^{\frac{-2\pi i(p-1)(j-1)}{N_x}}, \tag{3.21}$$

for $p = 2, \ldots, \lfloor \frac{N_x}{2} \rfloor + 1$. This means that $\mathbf{v}_{N_x-p+2}$ corresponds to both the (N_x-p+1)th and $(-p+1)$th wavenumber components of the Fourier series for $p = 2, \ldots, \lfloor \frac{N_x}{2} \rfloor + 1$. This cyclic property of the roots of unity is responsible for *aliasing* and will be discussed in Section 3.4. As $\mathbf{v}_{N_x-p+2}$ in (3.21) is complex conjugate to $\mathbf{v}_p$, despite its index being positive, it corresponds to the $(-p+1)$th Fourier basis function for $p = 2, \ldots, \lfloor \frac{N_x}{2} \rfloor + 1$. This is because the $(p-1)$th and $(-p+1)$th Fourier basis functions are complex conjugates in a Fourier series and the $(-p+1)$th Fourier basis function is resolvable on the finite grid [60] (see Section 3.4).

Since the numerical solution is constructed from a discrete sample of a finite number of wavenumber components from a Fourier series, we can relate the coefficients of a discrete Fourier series to those of a Fourier series, for the same function. Formulating the coefficients in this way will allow us to understand how the eigenvalues of the scheme propagate all wavenumber components of the Fourier series for the initial condition and how they introduce numerical dissipation and dispersion. We investigate this relationship in the next section, through considering the effects of *aliasing*.

3.4 Aliasing error

Aliasing errors are a form of *sampling error*. They occur due to the finite grid that the numerical method uses to solve the considered problem, only making use of a finite amount of information. Aliasing occurs when you only view a wavenumber component at the discrete mesh points of the grid and it appears to be a lower wavenumber component as a result. The lowest and highest resolvable wavenumber components of the Fourier series on the grid have wavenumber $k = -\frac{N_x}{2}$ and $k = \frac{N_x}{2}$ respectively [60] as these are constructed from Fourier basis functions which pass through every grid point of the domain. The real wavenumber component created by these wavenumber components has a wavenumber of $\frac{N_x}{2}$. This is known as the Nyquist rate [60, 67]. When N_x is even, $\mathbf{v}_{\frac{N_x}{2}+1}$ produces this wavenumber component. When N_x is odd, as N_x is not even, the wavenumber component corresponding to $\frac{N_x}{2}$ does not form a part of the numerical solution. The highest resolvable wavenumber component of the numerical solution in this case has real wavenumber $\frac{N_x-1}{2}$, constructed by $\mathbf{v}_{\frac{N_x+1}{2}}$ and $\mathbf{v}_{\frac{N_x+3}{2}}$.

In order to understand the effects of aliasing, consider the kth Fourier basis function of the Fourier series for $u(x,t)$, viewed at grid point x_{j-1},

$$e^{2\pi ikx_{j-1}} = e^{\frac{2\pi ik(j-1)}{N_x}} = \begin{cases} e^{2\pi i[k]_{N_x}x_{j-1}}, & \text{for } [k]_{N_x} = 0, \ldots, \lfloor \frac{N_x}{2} \rfloor, \\ e^{-2\pi i(N_x-[k]_{N_x})x_{j-1}}, & \text{for } [k]_{N_x} = \lfloor \frac{N_x}{2} \rfloor + 1, \ldots, N_x - 1, \end{cases}$$

where $j = 1, \ldots, N_x$. Here, $[\cdot]_{N_x}$ denotes modulo N_x. This means that any wavenumber component of the Fourier series with wavenumber $(p-1)+sN_x$, for $p = 1, \ldots, N_x$, $s \in \mathbb{Z}$, aliases to wavenumber $(p-1)$ when $p = 1, \ldots, \lfloor \frac{N_x}{2} \rfloor + 1$ and aliases to wavenumber $-(N_x - p + 1)$ when $p = \lfloor \frac{N_x}{2} \rfloor + 2, \ldots, N_x$. If in the latter case, if we substitute $q = N_x - p + 1$, we can see that these are the negative resolvable wavenumbers on the grid.

Due to the wavenumber index $(p-1)$ for $p = 1, \ldots, N_x$ of the 1D DFT basis, we will consider the wavenumber components of the Fourier series aliased onto these wavenumber components. However we must remember that when considering $p = \lfloor \frac{N_x}{2} \rfloor + 2, \ldots, N_x$, these wavenumber components in fact correspond to the negative resolvable wavenumber components on the finite grid. We will consider these wavenumber components when we refer to the resolvable wavenumber components of the solution.

Now we have explored the concept of aliasing, we can use it to determine how it impacts the construction of the coefficients for the 1D DFT basis, in comparison to the coefficients of the Fourier series, for the same function.

3.4.1 The Poisson summation

Consider the coefficients of the resolvable Fourier basis functions of $u(x, 0)$, as found through applying the 1D DFT to the function sampled at each grid point of the finite difference scheme,

$$\mathcal{F}_p(\mathbf{U}^0) = \mathbf{v}_p^* \mathbf{U}^0 = \frac{1}{\sqrt{N_x}} \sum_{j=1}^{N_x} u(x_{j-1}, 0) e^{\frac{-2\pi i(p-1)(j-1)}{N_x}},$$

for $p = 1, \ldots, N_x$. As $\mathbf{U}^0$ is created by sampling $u(x, 0)$, this discrete Fourier series is exact.

Representing the function $u(x, 0)$ by a Fourier series, allows the effects of aliasing on the coefficients of the resolvable wavenumber components to be seen through the *Poisson summation* [60]. In order to do this, the Fourier series needs to be equal to the function at all of the sample points. This requires the function to possess a convergent Fourier series and be continuous at all sample points in space. However, the Fourier series for the function from which the samples were taken, may not have these properties. As the 1D DFT of a function is not unique to that function [66], the Fourier series from an alternative function may be considered instead. This function must have a convergent Fourier series which when evaluated at the sample points in space, is equal to the original function. The alternative function need not be continuous at the sample points. It may possess a discontinuity at a sample point, whose midpoint is equal to the original function, at the sample point.

In order to demonstrate the Poisson summation, assume that the function $u(x, 0)$ is continuous and has a convergent Fourier series given by (3.6), at each grid point in

space. Then,

$$\mathcal{F}_p(\mathbf{U}^0) = \frac{1}{\sqrt{N_x}} \sum_{j=1}^{N_x} \sum_{k=-\infty}^{\infty} c_k e^{\frac{2\pi i k(j-1)}{N_x}} e^{\frac{-2\pi i(p-1)(j-1)}{N_x}} = \sqrt{N_x} \sum_{k=-\infty}^{\infty} c_{p-1+kN_x}, \quad (3.22)$$

for $p = 1, \ldots, N_x$. This is known as the Poisson summation [60]. Equation (3.22) shows that the coefficient of the $(p-1)$th resolvable wavenumber component found by the 1D DFT, is made up of the sum of the coefficients of the wavenumber components of the Fourier series, which are aliased to the $(p-1)$th wavenumber component. The coefficient of each unresolvable wavenumber component becomes a part of the coefficient for a resolvable wavenumber component. Applying the matrix M to the vector $\mathbf{U}^0$,

$$\mathcal{F}_p(\mathbf{U}^1) = \lambda_p \mathcal{F}_p(\mathbf{U}^0) = \sqrt{N_x} \sum_{k=-\infty}^{\infty} \lambda_p c_{p-1+kN_x}, \quad (3.23)$$

for $p = 1, \ldots, N_x$. This propagates the resolvable and unresolvable wavenumber components with wavenumber $(p-1+kN_x)$ for $k \in \mathbb{Z}$, using the eigenvalue λ_p, $p = 1, \ldots, N_x$. This allows M to propagate all the wavenumber components of the Fourier series, by only directly acting on N_x of them.

Another consequence of aliasing is *spectral leakage*. Spectral leakage occurs when a function is sampled over a non-integer multiple of its wavelength. Calculating the coefficients of the Fourier series for the function in this instance, results in the coefficients calculated containing contributions from wavenumber components, other than those already identified in (3.23) [60]. As we are sampling $u(x,t)$ over one complete period in this problem, there is no spectral leakage present.

Therefore the aliasing present in our considered problem, results in M applying the same magnitude and phase shift to an unresolvable wavenumber component, as it applies to the resolvable wavenumber component it aliases to. This is not necessarily the correct magnitude or phase shift for the considered unresolvable wavenumber component, even if it applies the correct magnitude and phase shift to the resolvable wavenumber component. This results in numerical dissipation and dispersion being introduced into the numerical solution.

3.5 Numerical dissipation and dispersion

Numerical dissipation and dispersion are important manifestations of numerical model error to consider, as their impact can be widespread and lead to physically unrealistic results. Limitations may sometimes be placed on model variables to avoid these effects, restricting the accuracy of the model [54].

Equation (3.23) shows that the eigenvalue $\lambda_{[k]_{N_x}+1}$, multiples the coefficient c_k of the Fourier series for the initial condition $u_0(x)$, to progress the state of the system forward Δt in time. We can define a one-periodic Fourier series for the numerical

solution, $w : \mathbb{R} \times [0, \infty) \to \mathbb{R}$ such that,

$$(x, t) \mapsto w(x, t) = \sum_{k=-\infty}^{\infty} v_k(t) e^{2\pi i k x} \tag{3.24}$$

where $v_k : [0, \infty) \to \mathbb{C}$,

$$t \mapsto v_k(t) = \lambda_{[k]_{N_x}+1}^{\frac{t}{\Delta t}} v_k(0), \quad v_k(0) = c_k, \quad \forall k \in \mathbb{Z}. \tag{3.25}$$

This Fourier series forms an approximation to the Fourier series for the analytical solution in (3.5). Evaluating (3.24) at a point that is not an integer multiple of Δx and/or Δt, interpolates the numerical solution in space and/or time, respectively. The definition of $v_k(0)$ results in $w(x, 0)$ being equal to the Fourier series of $u(x, 0)$ for all $x \in \mathbb{R}$.

Defining the function, $g_k^{scheme} : \mathbb{C} \to \mathbb{C}$ such that,

$$z \mapsto g_k^{scheme}(z) = \lambda_{[k]_{N_x}+1} z, \tag{3.26}$$

creates a function that maps the Fourier coefficients of (3.24) Δt through time. A similar function can be defined for (3.5) that maps its coefficients Δt through time. Define the function $g_k : \mathbb{C} \to \mathbb{C}$ such that $b_k(n\Delta t) \mapsto g_k(b_k(n\Delta t)) = b_k((n+1)\Delta t)$ for all $n \in \mathbb{N}_0$. Suppose that the functions $b_k(t)$ are invertible for all k, then $g_k(\cdot)$ is defined as,

$$g_k(\cdot) := b_k(b_k^{-1}(\cdot) + \Delta t). \tag{3.27}$$

The function $g_k^{scheme}(z)$ is an approximation to $g_k(z)$. The following demonstrates that $g_k^{scheme}(z)$ is defined similarly to $g_k(z)$ in (3.27), using the functions $v_k(t)$. Initially, note that,

$$v_k(t + \Delta t) = \lambda_{[k]_{N_x}+1} \lambda_{[k]_{N_x}+1}^{\frac{t}{\Delta t}} v_k(0) = \lambda_{[k]_{N_x}+1} v_k(t). \tag{3.28}$$

This then gives that if $v_k(\cdot)$ is invertible,

$$v_k(v_k^{-1}(z) + \Delta t) = \lambda_{[k]_{N_x}+1} v_k(v_k^{-1}(z)) = \lambda_{[k]_{N_x}+1} z = g_k^{scheme}(z). \tag{3.29}$$

The eigenvalues in (3.25) are scheme dependent, so each scheme may influence each c_k differently. In an ideal world, the functions $g_k^{scheme}(\cdot)$ and $g_k(\cdot)$ would be equal for all $k \in \mathbb{Z}$. This would result in $w(x, t)$ being equal to the Fourier series for $u(x, t)$. However, the Fourier series at time $t = 0$ has an infinite number of coefficients c_k, whilst the matrix implementing the scheme only possesses a finite number of eigenvalues. The eigenvalue λ_p is used to propagate the coefficients c_{p-1+kN_x} for $k \in \mathbb{Z}$, where $p = 1, \ldots, N_x$ as seen in (3.23). Tailoring λ_p to correctly propagate c_{p-1} for instance, does not necessarily mean that λ_p will correctly propagate c_{p-1+kN_x} for all $k \in \mathbb{Z} \setminus \{0\}$. There are a finite number of eigenvalues to tailor for an infinite number of Fourier

coefficients. As a result, an ideal finite difference scheme that correctly propagates all the Fourier coefficients is unlikely. In the next Section, we identify $g_k(\cdot)$ through the Fourier series solution to problem (3.1).

3.5.1 The Fourier series solution for the 1D linear advection problem

As we have already determined that it is the eigenvalues of our scheme which form $g_k^{scheme}(\cdot)$, we require the Fourier series for the analytical solution of problem (3.1), to determine $g_k(\cdot)$ for comparison. This function will also allow us to explore how the Fourier series for the analytical solution to problem (3.1), satisfies the properties of the solution, set out in Section 3.1.

In the case of problem (3.1), given the coefficients c_k for the Fourier series of $u_0(x)$, the Fourier series for $u(x,t)$ is given by,

$$u(x,t) \sim \sum_{k=-\infty}^{\infty} c_k e^{2\pi ik(x-\mu t)}. \tag{3.30}$$

As a result, $b_k(t) = c_k e^{-2\pi ik\mu t} = b_k(0)e^{-2\pi ik\mu t}$ and the time dependent portion of this coefficient is $e^{-2\pi ik\mu t}$. This makes $g_k(z) = e^{-2\pi ik\mu\Delta t}z$.

Therefore $g_k^{scheme}(z) = \lambda_{[k]_{N_x}+1}z$ is an approximation for $g_k(z) = e^{-2\pi ik\mu\Delta t}z$. The multiplying factor in these functions drive the changes in the Fourier coefficients in time Δt, so it is the magnitude and phase of these factors that are of interest when considering numerical dissipation and dispersion respectively. As a result, we are interested in how the magnitude and phase of $\lambda_{[k]_{N_x}+1}$ and $e^{-2\pi ik\mu\Delta t}$ compare.

The eigenvalue $\lambda_{[k]_{N_x}+1}$ is recovered from $g_k^{scheme}(z)$ by evaluating at $z = 1$. Similarly, $e^{-2\pi ik\mu\Delta t}$ is found by evaluating $g_k(z)$ at $z = 1$. Since the same change in time is applied by $g_k(\cdot)$ and $g_k^{scheme}(\cdot)$ with each application, only one application of both functions needs to be analysed. The complex conjugate property of the coefficients of Fourier series means that $g_k(z)$ and $g_{-k}(z)$ are complex conjugates for all $k \in \mathbb{Z}$ and $z \in \mathbb{C}$. The same is true for $g_k^{scheme}(z)$ and $g_{-k}^{scheme}(z)$. Therefore, we need only compare $g_k(1)$ and $g_k^{scheme}(1)$ for $k \in \mathbb{N}_0$. We now define numerical amplification, numerical dissipation and numerical dispersion through these quantities, based on the examples in [6].

Definition 3.3 (Numerical Amplification)**.** *Numerical amplification occurs when the magnitude of at least one wavenumber component of the numerical solution is increased during propagation. This means that there exists $k \in \mathbb{Z}$ such that $|\lambda_{[k]_{N_x}+1}| > |g_k(1)|$. This results in the growth of the corresponding real wavenumber component in the numerical solution, as time increases. The finite difference scheme is then termed numerically amplifying.*

Definition 3.4 (Numerical Dissipation)**.** *Numerical dissipation occurs when the magnitude of at least one wavenumber component of the numerical solution is decreased during propagation. This means that there exists $k \in \mathbb{Z}$ such that $|\lambda_{[k]_{N_x}+1}| < |g_k(1)|$. This results in the decay of the corresponding real wavenumber component in the numerical solution, as time increases. The finite difference scheme is then termed numerically dissipative.*

Definition 3.5 (Numerical Dispersion)**.** *Numerical dispersion occurs when the phase of at least one wavenumber component of the numerical solution is propagated incorrectly. This means that there exists $k \in \mathbb{Z}$ such that $e^{i\theta_{[k]_{N_x}+1}} \neq e^{iphase(g_k(1))}$. This results in the corresponding real wavenumber component progressing with the wrong speed in the numerical solution, as time increases. The finite difference scheme is then termed numerically dispersive.*

Definitions 3.4 and 3.5 represent a formalisation of the definitions for numerical dissipation and dispersion presented through example in Durran [6, p. 49]. Definition 3.3 is defined to complement the definition of numerical dissipation and to emphasise the idea that studying the effects of numerical amplification is just as important as studying the effects of numerical dissipation. Schemes containing numerical amplification effects are described as numerically unstable, however by studying the numerical amplification properties of some schemes, it may be possible to modify them to compensate for these effects. Williams [68] was able to analyse the numerically amplifying and dissipative properties of the Leapfrog scheme to improve the Robert-Asselin filter. In this thesis, the considered finite difference schemes will only be considered when the schemes are numerically stable, so the only numerical errors they can introduce are numerical dissipation and/or numerical dispersion. We now analyse $g_k(1)$ for the 1D linear advection problem, with respect to these definitions.

The magnitude of $g_k(1) = e^{-2\pi ik\mu\Delta t}$ is one for all k, which means that the amplitude of each wavenumber component remains unchanged over time. In the case of this particular problem, if all the eigenvalues of the scheme have unit magnitude, then all the wavenumber components of the Fourier series are propagated with the correct magnitude over time and the finite difference scheme is termed numerically non-dissipative with respect to all wavenumber components. In this case, any aliasing errors in the numerical solution are not due to numerical dissipation.

A finite difference scheme is numerically non-dispersive when $e^{i\theta_{[k]_{N_x}+1}} = e^{-2\pi ik\mu\Delta t}$ for all $k \in \mathbb{Z}$. However, as mentioned previously, this is extremely hard to accomplish due to there only being N_x eigenvalues. The following Remark aims to identify if it is possible to define the eigenvalues of a finite difference scheme, such that they correctly

propagate N_x wavenumber components of the numerical solution, but choose parameters within these eigenvalues so that all wavenumber components of the numerical solution are correctly propagated.

Remark 3.6. *Suppose* $\lambda_p = g_{p-1+sN_x}(1) = e^{-2\pi i(p-1+sN_x)\mu\Delta t}$ *for some* $s \in \mathbb{Z}$ *and* $p = 1, \ldots, N_x$. *Then* λ_p *will correctly propagate the* $(p-1+sN_x)$*th wavenumber component of the solution. The Poisson summation gives that this eigenvalue propagates the wavenumber components corresponding to wavenumbers* $(p-1+kN_x)$ *for all* $k \in \mathbb{Z}$. *Examining the magnitude of* λ_p, *we find that* $|\lambda_p| = 1$. *This is the correct magnitude for all the wavenumber components that the eigenvalue propagates. Examining the phase of* λ_p *we obtain,*

$$\lambda_p = e^{-2\pi i(p-1+sN_x)\mu\Delta t} = e^{-2\pi i(p-1+kN_x)\mu\Delta t} e^{-2\pi i(s-k)h}, \tag{3.31}$$

where $h = \frac{|\mu|\Delta t}{\Delta x} = |\mu|\Delta t N_x$ *is the CFL number. Then the phase of* λ_p *is correct for propagating the* $(p-1+kN_x)$*th wavenumber component for some* $k \in \mathbb{Z}$, *when* $(s-k)h = \alpha$, *for some* $\alpha \in \mathbb{Z}$.

We will consider h *to be a rational constant as its constituent variables are real and in numerical simulations, we are only able to define them using rational values. Let* $h = \frac{q}{b}$ *such that* $q, b \in \mathbb{N}$ *and* $gcd(q, b) = 1$ *(greatest common divisor), then we require* $\frac{(s-k)q}{b} = \alpha \in \mathbb{Z}$. *If* $k = s$ *then the corresponding wavenumber component is correctly propagated. Now consider* $k \neq s$, *this requires that* $b|(s-k)$ *for all* $k \in \mathbb{Z} \setminus \{s\}$ *as* $b \nmid q$. *Here* $\cdot|\cdot$ *denotes divides ie:* $b|(s-k)$ *is equivalent to* $\frac{s-k}{b} \in \mathbb{Z}$. *In order for* b *to divide* $(s-k)$ *for all* $k \in \mathbb{Z} \setminus \{s\}$, *we require* $b = 1$, *therefore* $h \in \mathbb{N}$.

So tailoring λ_p *to propagate the* $(p-1+sN_x)$*th wavenumber component, results in the amplitude of the wavenumber components with wavenumbers* $(p-1+kN_x)$, *being correctly propagated for all* $k \in \mathbb{Z}$. *However, the phase of these wavenumber components is only propagated correctly for all* k *when* $h \in \mathbb{N}$.

Remark 3.6 shows that for problem (3.1), when a numerically stable finite difference scheme is numerically non-dissipative with respect to the resolvable wavenumber components, it is numerically non-dissipative with respect to all wavenumber components of the numerical solution. It also shows that when a scheme is numerically non-dispersive with respect to the resolvable wavenumber components of the solution, it is not non-dispersive with respect to all wavenumber components of the numerical solution, unless $h \in \mathbb{N}$.

As the eigenvalues of the scheme are defined to propagate the resolvable wavenumber components of the solution, the numerically dissipative and/or dispersive properties of the scheme can initially be discussed with respect to the resolvable wavenumber components of the solution. The numerically dissipative and/or dispersive effects of these resolvable wavenumber components on the unresolvable wavenumber components are a consequence of aliasing and will be discussed as such.

The analysis of $g_k(1) = e^{-2\pi ik\mu\Delta t}$ in this Section has revealed that for problem (3.1), if a numerically stable scheme is:

- numerically non-dissipative and non-dispersive with respect to the resolvable wavenumber components, aliasing will still occur if $h \in \mathbb{R}^+ \setminus \mathbb{N}$, but this will be a form of numerical dispersion (MNIMC scheme - see Section 3.7.3),
- numerically dissipative and non-dispersive with respect to the resolvable wavenumber components, aliasing will be a form of numerical dissipation, but will also be a form of numerical dispersion if $h \in \mathbb{R}^+ \setminus \mathbb{N}$ (Upwind scheme for $h = 0.5$),
- numerically non-dissipative and dispersive with respect to the resolvable wavenumber components, aliasing will be a form of numerical dispersion (Preissman Box scheme for $0 < h < 1$),
- numerically dissipative and dispersive with respect to the resolvable wavenumber components, aliasing will be a form of numerical dissipation and dispersion (Lax-Wendroff scheme for $0 < h < 1$).

The analysis leading to the development of the functions $g_k(z)$ and $g_k^{scheme}(z)$ and hence Definitions 3.4 and 3.5, was based on the examples of numerical dissipation and dispersion found in Durran [6] and Vreugdenhil [69]. These authors provide an alternative method for determining the numerically dissipative and dispersive properties of finite difference schemes. These are the *damping factor* and *relative phase* found in the following Sections.

3.5.2 The damping factor

In order to characterise numerical dissipation, Vreugdenhil [69] proposes measuring the *damping factor*. This is defined using the variables of this thesis as follows,

$$A_k := \left| \frac{g_k^{scheme}(1)}{g_k(1)} \right| = \left| \frac{\lambda_{[k]_{N_x}+1}}{g_k(1)} \right|, \quad g_k(1) \neq 0, \quad k \in \mathbb{Z}. \tag{3.32}$$

If $A_k = 1$ for all $k \in \mathbb{Z}$, then the finite difference scheme propagates the magnitude of the coefficients c_k correctly. Hence the scheme is numerically non-dissipative. If there exists $k \in \mathbb{Z}$ such that $A_k > 1$, or $A_k < 1$, the scheme is numerically amplifying or numerically dissipative respectively [6]. However, if $g_k(1) = 0$, A_k cannot be calculated.

3.5.3 The relative phase

In order to characterise numerical dispersion, Vreugdenhil [69] and Durran [6] both propose measuring the *relative phase*. This is defined using the variables of this thesis as follows,

$$R_k := \frac{\text{phase}(g_k^{scheme}(1))}{\text{phase}(g_k(1))} = \frac{\theta_{[k]_{N_x}+1}}{\text{phase}(g_k(1))}, \quad \text{phase}(g_k(1)) \neq 0, \quad k \in \mathbb{Z}. \tag{3.33}$$

If $R_k = 1$ for all k, then the finite difference scheme is not introducing any phase errors into the coefficients c_k. Hence, the scheme is numerically non-dispersive. If there exists $k \in \mathbb{Z}$ such that $R_k > 1$ or $R_k < 1$, the scheme is numerically dispersive [6]. If the phase of $g_k(1)$ is zero, then this cannot be calculated. There is also the problem that if the phase of $g_k(1)$ and $\theta_{[k]_{N_x}+1}$ are not in the same 2π period, we may find that for some k where $e^{i\text{phase}(g_k(1))} = e^{i\theta_{[k]_{N_x}+1}}$, R_k is not equal to one. This means that despite the kth wavenumber component being numerically non-dispersive, the relative phase indicates that it is numerically dispersive.

3.6 Analysis of finite difference schemes for the 1D linear advection problem

In order to use the considered schemes to solve problem (3.1) numerically, the schemes need to converge to the analytical solution of the problem [70]. A finite difference scheme will converge to the true solution if it is both consistent and numerically stable, by the Lax-Richtmyer Equivalence Theorem [14, 71]. Hence it is important to understand when these properties hold for all our considered schemes. These properties are presented in Table 3.1.

We are also interested in the numerically dissipative and dispersive properties of our schemes, determined by the eigenvalues of the scheme, when the schemes are numerically stable. Initially we can investigate the numerically dissipative and dispersive properties of the schemes, with respect to the resolvable wavenumber components they correspond to and then relate the analysis to the unresolvable wavenumber components of the numerical solution. We choose to identify the numerically dissipative and dispersive properties of our considered finite difference schemes, by directly examining the eigenvalues of the schemes.

Section 6.4.3 has shown that a scheme for solving problem (3.1), is numerically non-dissipative when $|\lambda_p| = 1$, for all $p = 1, \ldots, N_x$. Therefore the numerically dissipative properties of the scheme are found by determining when the eigenvalues have unit magnitude. Section 3.5.1 also showed that a scheme solving problem (3.1) that has eigenvalues with a linear phase with respect to wavenumber, have the potential to be numerically non-dispersive with respect to the resolvable wavenumber components. Therefore the numerically dispersive properties can be investigated by determining when the phase of the eigenvalues are linear with respect to wavenumber. This can be achieved by determining the phase of the eigenvalues and considering $z = \frac{p-1}{N_x}$ as a continuous variable. The phase can then be differentiated to identify when it is a linear function with respect to z. These methods are used to determine the numerically dissipative and dispersive properties of the schemes in Table 3.2.

Another method that could be used to identify the numerically dissipative and dispersive properties of a finite difference scheme, is to analyse the *modified equation*

[6] associated with it. The modified equation approach was developed based upon the idea that, if the considered finite difference scheme is introducing numerical model error, it must be producing the solution to a slightly different PDE. Taylor expansions are used to develop this modified PDE, resulting in the limitations on when Taylor expansions are valid also being placed on the validity of the modified PDE. Therefore, we choose not to use the modified PDE approach due to these limitations.

3.6.1 The Upwind scheme

The Upwind scheme in (3.10), is an explicit finite difference scheme derived to solve the 1D linear advection problem, by approximating the temporal derivative using a forward difference in time and the spatial derivative by a backward difference in space. The scheme is only numerically stable for $\mu > 0$, as the backwards difference in space results in the scheme propagating information in the positive x-direction over time. When $\mu < 0$, the *downwind* scheme is used instead. This uses a forwards difference to approximate the spatial derivative, resulting in information propagating in the negative x-direction over time [72].

The eigenvalues of the matrix implementing the Upwind scheme are,

$$\lambda_p = 1 + h\left\{\cos\left[\frac{2\pi(p-1)}{N_x}\right] - 1\right\} - ih\sin\left[\frac{2\pi(p-1)}{N_x}\right], \tag{3.34}$$

for $p = 1, \ldots, N_x$.

3.6.2 The Preissman Box scheme

The Preissman Box scheme in (3.11) is an implicit finite difference scheme derived to solve the 1D linear advection problem, by approximating the temporal derivative by the average of two forward differences in time, one taken at x_j and the other at x_{j-1}. The spatial derivative is approximated similarly by the average of two forward differences in space, one taken at t^{n+1} and the other at t^n [64].

The eigenvalues of the matrix implementing the Preissman Box scheme are,

$$\lambda_p = \frac{(1-h^2) + (1+h^2)\cos\left[\frac{2\pi(p-1)}{N_x}\right] - 2ih\sin\left[\frac{2\pi(p-1)}{N_x}\right]}{(1+h^2) + (1-h^2)\cos\left[\frac{2\pi(p-1)}{N_x}\right]}, \tag{3.35}$$

for $p = 1, \ldots, N_x$. We note that $|\lambda_p| = 1$ for all $p = 1, \ldots, N_x$.

3.6.3 The Lax-Wendroff scheme

The Lax-Wendroff scheme in (3.12) is an explicit finite difference scheme derived to solve the 1D linear advection, by Taylor expanding $u(x_j, t^{n+1})$ about (x_j, t^n) in terms of spatial derivatives, truncating after the 2nd order term. The temporal derivatives

of the expansion are then approximated by central differences in time and the spatial derivative by a backward difference in space [59].

The eigenvalues of the matrix implementing the Lax-Wendroff scheme are,

$$\lambda_p = 1 + h^2 \left\{ \cos\left[\frac{2\pi(p-1)}{N_x}\right] - 1 \right\} - ih \sin\left[\frac{2\pi(p-1)}{N_x}\right], \tag{3.36}$$

for $p = 1, \ldots, N_x$.

3.6.4 Finite difference scheme property summary

Tables 3.1 and 3.2 summarise the properties of the Upwind, Preissman Box and Lax-Wendroff schemes as identified through the eigenvalues of the schemes in Sections 3.6.1-3.6.3. In particular, the numerically dissipative and/or dispersive properties were derived when considering the schemes to be numerically stable. The properties of the NIMC and MNIMC schemes described in Sections 3.7.1 and 3.7.3 respectively, are also included in the Tables for comparison. The schemes are all identical when $h = 1$, forming a scheme which is numerically non-dissipative and non-dispersive with respect to all wavenumber components of the numerical solution.

The aim of this thesis was to determine the effects of numerical model error on the analysis vector produced through strong constraint 4D-Var data assimilation. We have chosen to investigate this through problem (3.1), which can be solved numerically by any of the schemes in Tables 3.1 and 3.2. Our reasons for choosing to use the Upwind, Preissman Box and Lax-Wendroff schemes, can be seen by examining the numerically dissipative and dispersive properties of the schemes in Table 3.2, for $h = 0.5$.

When $h = 0.5$, the Upwind scheme is numerically dissipative and non-dispersive with respect to the resolvable wavenumber components of the solution, the Preissman Box scheme is numerically non-dissipative and dispersive with respect to the resolvable wavenumber components of the solution, the Lax-Wendroff is both numerically dissipative and dispersive with respect to the resolvable wavenumber components of the solution and the MNIMC scheme is numerically non-dissipative and non-dispersive with respect to all resolvable wavenumber components. As a result, the effects of numerical dissipation and dispersion on the analysis vector can be investigated as individual forms of error (Upwind and Preissman Box schemes), in unison (Lax-Wendroff scheme) and as aliasing errors (MNIMC scheme), with respect to the resolvable wavenumber components of the solution. Figure 3.1 presents the properties of the eigenvalues of each scheme, when $h = 0.5$. For simplicity, all numerical results will be generated for the 1D linear advection problem using $\mu = 1$.

As discussed in Section 3.3.1, $\overline{\lambda_p} = \lambda_{N_x-p+2}$, for $p = 2, \ldots, N_x$. Therefore λ_p and λ_{N_x-p+2} are complex conjugates for $p = 2, \ldots, \frac{N_x+1}{2}$ (as N_x is odd in Figure 3.1). We remind the reader that the eigenvalues λ_p for $p = 1, \ldots, \frac{N_x+1}{2}$ and $p = \frac{N_x+3}{2}, \ldots, N_x$ correspond to wavenumber components with wavenumber $k = p-1$ and $k = -N_x+p-1$

respectively, in a Fourier series. Therefore only $p = 1, \ldots, \frac{N_x+1}{2}$ need be examined to determine the effect of the eigenvalues on real wavenumber components $k = p-1$. The Nyquist rate lies between $p = \frac{N_x+1}{2}$ and $p = \frac{N_x+3}{2}$ at the mid-points of the horizontal axis in Figures 3.1(a) and 3.1(b). The Nyquist rate is responsible for the vertical line of symmetry through the point $\frac{p-1}{N_x} = 0.5$ on the horizontal axis of Figure 3.1(a) and the rotational symmetry of Figure 3.1(b) about the point $(0, 0.5)$.

The number of discretisation points N_x, is chosen to be odd in Figure 3.1 as this is a requirement for the MNIMC scheme (see Section 3.7.3). This condition is also required for the matrix implementing the Upwind scheme to be invertible when $h = 0.5$. As we will see later on, we do not need this to be true to perform our analysis using the Upwind scheme.

Figure 3.1(a) shows the magnitude of the eigenvalues in the spectrum of the considered finite difference schemes. The line of symmetry in the plot between $\frac{p-1}{N_x} = \frac{N_x-1}{2N_x} = \frac{50}{101}$ and $\frac{p-1}{N_x} = \frac{N_x+1}{2N_x} = \frac{51}{101}$ shows the complex conjugacy of λ_p and λ_{N_x-p+2} for $p = 2, \ldots, \frac{N_x+1}{2}$. Both the Preissman Box and MNIMC scheme are shown to be numerically non-dissipative with respect to the resolvable wavenumber components of the solution. The Upwind and Lax-Wendroff schemes are both numerically dissipative with respect to the resolvable wavenumber components of the scheme. As wavenumber increases to the Nyquist rate (the point of the line of symmetry), the attenuation effects of the schemes increase, with the Upwind scheme attenuating its wavenumber components more.

Figure 3.1(b) plots the phase of the eigenvalues in the spectrum of the considered finite difference schemes. The plot demonstrates the complex conjugacy property of the eigenvalues by showing that $\theta_p = -\theta_{N_x-p+2}$ for $p = 2, \ldots, \frac{N_x+1}{2}$, creating rotational symmetry in the plot. The Upwind and MNIMC schemes are shown to be numerically non-dispersive with respect to the resolvable wavenumber components as the phase is linear with respect to $\frac{p-1}{N_x}$. The Preissman Box and Lax-Wendroff schemes are both numerically dispersive with respect to the resolvable wavenumber components of the solution. Examining the phase of the Preissman Box and Lax-Wendroff schemes for $p = 1, \ldots, \frac{N_x+1}{2}$, we can see that when $p > 1$, the resolvable real wavenumber components are propagated too fast and too slow respectively.

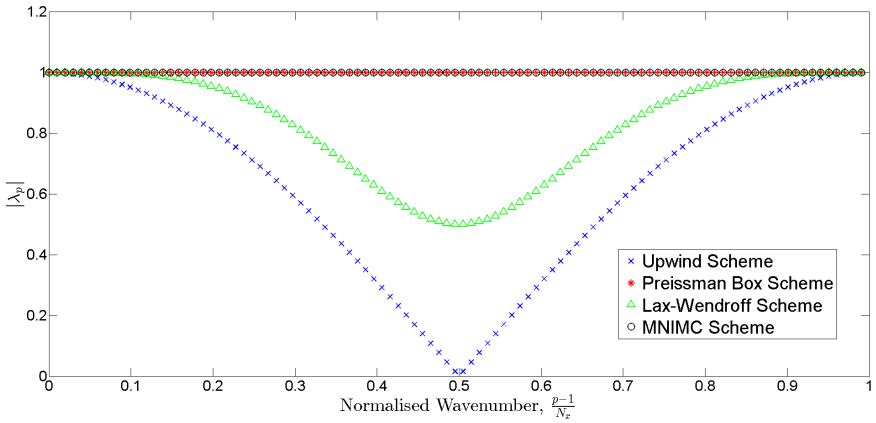
Figure 3.1(c) displays the properties shown by Figures 3.1(a) and 3.1(b), in the form of an argand diagram for the eigenvalues of the spectrum.

Scheme	Consistent	Numerically Stable	Convergent	Singular Matrix
Upwind	Always	$0 < h \leq 1$	$0 < h \leq 1$	N_x is even and $h = 0.5$
Preissman Box	Always	Always	Always	Never
Lax-Wendroff	Always	$0 < h \leq 1$	$0 < h \leq 1$	N_x is even and $h = \frac{1}{\sqrt{2}}$
NIMC	$h = 1$	$0 < h \leq 1$	$h = 1$	Never
MNIMC	Always	Always	Always	Never

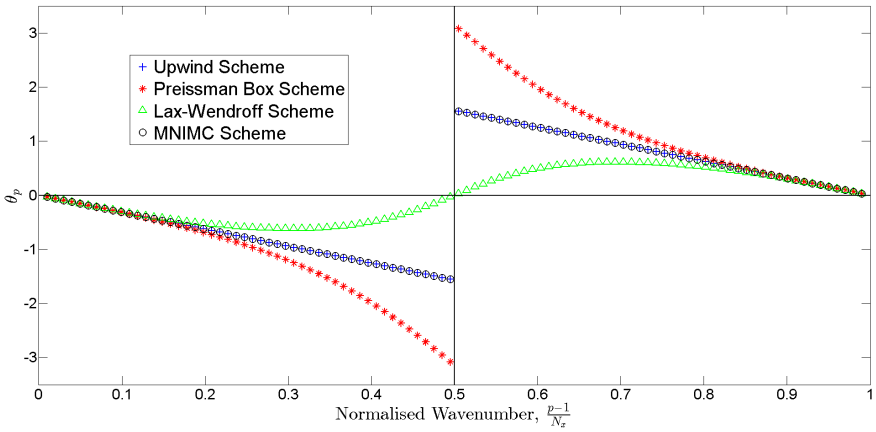
Table 3.1: *This Table summarises the consistency, numerical stability and hence convergence properties, for the finite difference schemes considered for solving problem (3.1). The consistency of the scheme is for sufficiently smooth initial conditions. Information on the invertibility of the matrix used to implement the scheme is also provided.*

Scheme	Non-Dissipative wrt resolvable wavenumber components	Non-Dispersive wrt resolvable wavenumber components	Non-Dissipative wrt all wavenumber components	Non-Dispersive wrt all wavenumber components
Upwind	$h = 1$	$h = 1$ or $h = 0.5$	$h = 1$	$h = 1$
Preissman Box	Always	$h = 1$	Always	$h = 1$
Lax-Wendroff	$h = 1$	$h = 1$	$h = 1$	$h = 1$
NIMC	$h = 1$	$h = 1$	$h = 1$	$h = 1$
MNIMC	Always	Always	Always	$h \in \mathbb{N}$

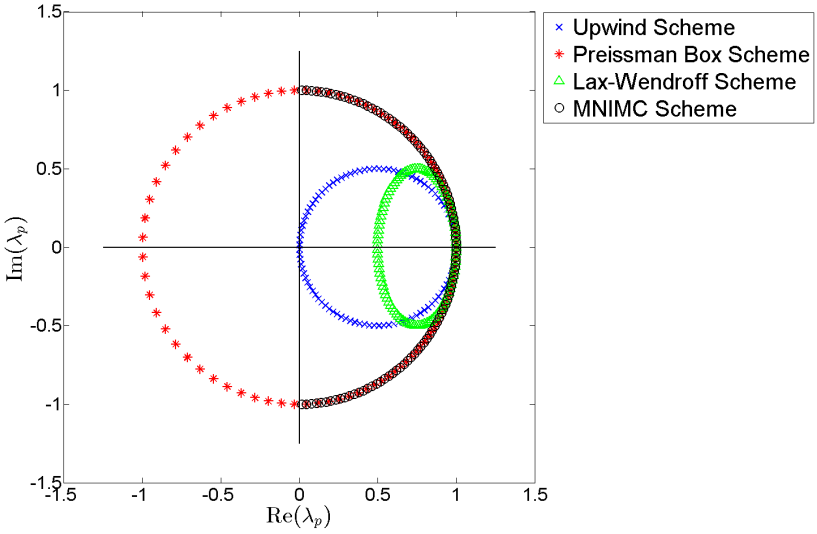
Table 3.2: *This Table summarises the numerically dissipative and dispersive properties with respect to the resolvable wavenumber components and all wavenumber components of the numerical solution, for the finite difference schemes considered for solving problem (3.1), for* $0 < h \leq 1$. *Here 'wrt' denotes 'with respect to'.*



(a) A plot of the magnitude of the eigenvalues in the spectrum of each considered finite difference scheme, given by $|\lambda_p|$ for $p = 1, \ldots, N_x$.



(b) A plot of the phase of the eigenvalues in the spectrum of each considered finite difference scheme, given by θ_p for $p = 1, \ldots, N_x$, $\theta_p \in [-\pi, \pi)$.



(c) An argand diagram for the eigenvalues in the spectrum of each considered finite difference scheme, given by λ_p for $p = 1, \ldots, N_x$.

Figure 3.1: *These plots demonstrate the numerically dissipative and dispersive properties of the Upwind, Preissman Box, Lax-Wendroff and MNIMC schemes when* $h = 0.5$, $\mu = 1$ *and* $N_x = 101$ *(*$\Delta t = \frac{1}{202}$*).*

3.6.5 The CFL condition

Examining Tables 3.1 and 3.2, we see that the CFL number h for problem (3.1), determines the convergence of the explicit schemes [7, 13] and the numerically dissipative and dispersive properties of each finite difference scheme. The CFL number for the 1D linear advection problem is defined as $h = \frac{|\mu|\Delta t}{\Delta x}$ and is named as such due to the CFL condition [70]. The CFL condition named after its inventors Courant, Friedrichs and Lewy, is a necessary condition for the convergence of explicit finite difference approximations [7]. In our considered problem, we have chosen $\mu > 0$, so $h = \frac{\mu\Delta t}{\Delta x}$. The CFL condition arises from the requirement that the domain of dependence of the PDE be contained within the domain of dependence of the finite difference scheme. A derivation can be found in [14, 73]. The domain of dependence for the 1D linear advection problem lies within the domain of dependence for each of our schemes when $h \leq 1$ [14, 70]. Hence, this is the CFL condition for our considered problem. Examining Table 3.1, we can see that the CFL condition is a necessary and sufficient condition for the convergence of both the Upwind and Lax-Wendroff schemes. The CFL number changes with the dimension of the considered system in space. We derive the CFL condition for the 2D linear advection problem in Section 5.5.4.

It is important to note that μ features in the CFL number for the 1D linear advection problem, as μ is the speed of propagation along the characteristic solution [6]. It is also the phase, group and wave speed for the problem. This is not usually the case, so it is important to remember that μ's presence in the CFL number, is due to the characteristic speed and not the phase, group or wave speed of the analytical solution.

Consider the Upwind and Lax-Wendroff schemes and suppose we choose Δt and Δx such that the CFL condition holds. Then keeping the ratio $\frac{\Delta t}{\Delta x}$ constant whilst letting $\Delta t, \Delta x \to 0$, will ensure that the domain of dependence of the finite difference scheme remains fixed. This will become important when we investigate orders of convergence in Chapter 4. The CFL condition is also a necessary condition for the numerical stability of finite difference schemes [7, 14].

Section 3.6.4 has demonstrated how important the CFL number is in determining the properties of each of our considered schemes. This includes when the scheme can be used to produce a numerical approximation to the solution of problem (3.1) and the type of error that is present in this approximation. Whilst investigating our considered schemes, a fixed CFL number will be chosen for each scheme. This will ensure that properties of the scheme are not changed, allowing the schemes to produce a numerical approximation to the true solution of problem (3.1) and for the numerically dissipative and dispersive properties of the schemes to remain unchanged. This will allow for the effects of particular numerically dissipative and dispersive properties of schemes, on the results of strong constraint 4D-Var data assimilation, to be investigated.

3.7 Generating perfect observations

In order to identify the effects of numerical model error in strong constraint 4D-Var data assimilation, we have specified that we require perfect observations of the physical system. Since problem (3.1) forms our physical system, we need a way to generate perfect observations of the system for numerical experiments and a way to construct them algebraically for the purposes of our analysis.

We have an analytic expression for the solution of problem (3.1) in the form of $u_0([x-\mu t]_1)$, however this form is not convenient for analytically calculating the error in the analysis vector arising from a scheme which has the form $\mathbf{U}^n = M^n\mathbf{U}^0$. It would be convenient for comparison analytically if perfect observations could be formulated in part by a numerically non-dissipative and non-dispersive finite difference scheme, with respect to the resolvable wavenumber components of the numerical solution. This would allow the effects of numerical dissipation and dispersion in the resolvable wavenumber components to be isolated from the affects of aliasing. The analytical solution $u_0([x-\mu t]_1)$ is suitable for calculating perfect observations numerically, however lots of function evaluations is computational expensive for large N_x. The Fourier series form of the analytical solution is convenient for analytically representing perfect observations and will be used in the following, but cannot be used to generate perfect observations numerically. The following sections identify ways for generating perfect observations using finite difference schemes.

3.7.1 The NIMC scheme

Consider the following finite difference scheme, the *Numerical Implementation of the Method of Characteristics* (NIMC),

$$U_j^{n+1} = \text{sgn}(\mu)hU_{j-1}^n.$$

This is an explicit finite difference scheme, which can be implemented similarly to the schemes considered in Section 3.3, via a circulant matrix $M_{NIMC} \in \mathbb{R}^{N_x\times N_x}$, under the conditions in Assumptions 3.2. When $h = 1$, the scheme takes the state of the system at (x_j, t^n) and moves it Δx in the positive direction along the x-axis in time Δt, to (x_{j+1}, t^{n+1}) when $\mu > 0$ as required of the solution to problem (3.1). The scheme follows the characteristic equation through the point (x_j, t^{n+1}). Similarly in the negative direction along the x-axis in time Δt, if $\mu < 0$. Examining Table 3.1, we see that the scheme is only convergent when $h = 1$. Therefore we can only generate a numerical solution from the scheme under this condition. Table 3.2 does tell us that this solution is exact as the scheme is numerically non-dissipative and non-dispersive with respect to all wavenumber components of the numerical solution when $h = 1$.

3.7.2 Problems with generating perfect observations using the NIMC scheme

The NIMC scheme looks like a promising scheme for providing perfect observations for our analysis of the effects of numerical model error on strong constraint 4D-Var data assimilation. We therefore need to investigate whether it can produce observations at the same points in time and space as the Upwind, Preissman Box and Lax-Wendroff scheme, despite being limited to $h = 1$.

The Upwind, Preissman Box, Lax-Wendroff and NIMC schemes all generate numerical solutions to problem (3.1) at each grid point in the domain, every Δt in time, where $\Delta t = \frac{h\Delta x}{|\mu|}$. We are only considering the Upwind and Lax-Wendroff schemes when they are numerically stable, so by Table 3.1 we limit the CFL number to $0 < h \leq 1$ for our analysis. Let us fix Δx. The wave speed $\mu > 0$ is fixed by problem (3.1). Now consider $h = 1$ for the NIMC scheme and $h < 1$ for the Upwind, Preissman Box and Lax-Wendroff schemes. Then the Upwind, Preissman Box and Lax-Wendroff schemes generate the state of the system over shorter time intervals than the NIMC scheme. Observations generated by the NIMC scheme could be interpolated in time, but this would add an additional error into the problem. An alternative method is required to generate perfect observations at the same points in space and time as the Upwind, Preissman Box and Lax-Wendroff schemes.

There are several options to solve this particular problem. The first is to use the NIMC scheme to generate perfect observations, by adjusting the value of Δx used in the scheme. The NIMC scheme generates observations every $\Delta t^{NIMC} = \frac{\Delta x^{NIMC}}{|\mu|}$ using $h = 1$, but we require $\Delta t = \frac{h\Delta x}{|\mu|}$ where h and Δx correspond to either the Upwind, Preissman Box or Lax-Wendroff scheme. In order to create perfect observations every Δt from both schemes, Δx^{NIMC} for the NIMC scheme is chosen to be equal to $h\Delta x$. We again suppose that h is a rational number of the form $h = \frac{q}{b}$ where $q, b \in \mathbb{N}$ and $\gcd(q, b) = 1$. Then choosing $\Delta x^{NIMC} = h\Delta x$ is equivalent to dividing Δx into b pieces and applying the NIMC scheme q-times. This is equivalent to increasing the speed of the numerical solution by a factor of q. This scheme is implemented by an $N_x b \times N_x b$ matrix, increasing the computational resources required to create perfect observations and complicating any algebraic analysis involving these perfect observations.

An alternative solution to the problem of numerically generating perfect observations is available for the linear advection problem. As the true solution to problem (3.1) preserves the shape of the solution and shifts it in the positive direction along the x-axis with constant speed μ, the *circshift* function in MATLAB®[74] can be used to shift the initial condition $|\mu|\Delta t = \frac{h}{N_x} = h\Delta x$ in space, to create perfect observations. This can be achieved similarly to the previous option by dividing Δx into b equally spaced pieces and applying *circshift* [74] so that it moves q of these new grid points in time Δt. Every b grid points of the new grid match up with the old grid points. These points can be used to create the relevant observations. This requires that the initial

condition be known every $\frac{\Delta x}{b}$ in space. This option also increases the computational cost by requiring extra discretisation points in space. However, this is significantly less than for the previous option and is easier to use algebraically.

When this approach is not feasible, an alternative method is required, one that can be easily implemented and analysed algebraically. A new finite difference scheme implemented by a real $N_x \times N_x$ matrix, could be defined. This scheme would be numerically non-dissipative and non-dispersive with respect to the resolvable wavenumber components of the solution, for any chosen value of the CFL number h, such that the scheme is consistent, convergent and numerically stable. In particular, we require such a scheme when $0 < h \leq 1$, so the Upwind and Lax-Wendroff scheme are convergent. In the case of the 1D linear advection problem, creating a scheme that is numerically non-dissipative with respect to the resolvable wavenumber components, creates a scheme that is numerically non-dissipative with respect to all wavenumber components of the solution. It is unlikely that we would be able to construct a scheme that is always numerically non-dispersive with respect to all wavenumber components, however it may be possible to define one that is numerically non-dispersive with respect to all resolvable wavenumber components. The following Section aims to develop such a scheme.

3.7.3 The MNIMC scheme

Consider the NIMC scheme for $h = 1$. At this point, the NIMC scheme is numerically non-dissipative and non-dispersive with respect to all wavenumber components. This scheme could potentially be modified to attain our goal of achieving a finite difference scheme that is numerically non-dissipative and non-dispersive with respect to all resolvable wavenumber components of the solution, for any value of h, but in particular when $0 < h \leq 1$. Let $\left\{\tilde{\lambda}_p\right\}_{p=1}^{N_x}$ denote the set of eigenvalues for the scheme we wish to derive. Given any value of h, the desired scheme needs to satisfy the following properties,

- be consistent, convergent and numerically stable,
- $\tilde{\lambda}_p^{\frac{1}{h}} = e^{\frac{-2\pi i(p-1)\mathrm{sgn}(\mu)}{N_x}}$ for all $p = 1, \ldots, N_x$,
- $\tilde{\lambda}_1 \in \mathbb{R}$ and $\overline{\tilde{\lambda}_p} = \tilde{\lambda}_{N_x-p+2}$ for $p = 2, \ldots, N_x$,
- eigenvectors given by the 1D DFT basis,
- $|\tilde{\lambda}_p| = 1$ for all $p = 1, \ldots, N_x$ ie: numerically non-dissipative with respect to all wavenumber components of the numerical solution,
- numerically non-dispersive with respect to all resolvable wavenumber components of the numerical solution.

The condition that $\tilde{\lambda}_p^{\frac{1}{h}} = e^{\frac{-2\pi i(p-1)\mathrm{sgn}(\mu)}{N_x}}$ for all $p = 1, \ldots, N_x$, can be explained as follows. The CFL number gives us that $\Delta t = \frac{h\Delta x}{|\mu|}$. Suppose we apply the new scheme $\frac{1}{h}$ times. Then the time that has passed is $\frac{\Delta t}{h} = \frac{\Delta x}{|\mu|}$. Suppose we run the NIMC scheme

($h_{NIMC} = 1$) using the same Δx. Then $\Delta t_{NIMC} = \frac{\Delta x}{|\mu|} = \frac{\Delta t}{h}$, by rearranging $h = \frac{|\mu|\Delta t}{\Delta x}$. This implies that if we apply our new scheme $\frac{1}{h}$ times, we aim to recover the eigenvalues of the NIMC scheme, as they are numerically non-dissipative and non-dispersive with respect to all wavenumber components of the numerical solution. Hence the second condition in the list.

Initially consider the condition that $\tilde{\lambda}_p^{\frac{1}{h}} = e^{\frac{-2\pi i(p-1)\text{sgn}(\mu)}{N_x}}$ and use this to give trial eigenvalues for the scheme, $\tilde{\lambda}_p = e^{\frac{-2\pi i(p-1)\text{sgn}(\mu)h}{N_x}}$ for all $p = 1, \ldots, N_x$. These eigenvalues do not satisfy the complex conjugate condition. We also notice that if N_x is even $\tilde{\lambda}_{\frac{N_x}{2}+1}$ is complex, rather than a real number. This will be a problem as its corresponding eigenvector in the 1D DFT basis is real, so requires a real eigenvalue to propagate it so that the result of the scheme remains real. Any solution produced by the scheme when N_x is even, would be complex. As a result, we cannot define this scheme for even N_x and *restrict ourselves to odd* N_x. We also modify the eigenvalues so that this trial form is retained for $p = 1, \ldots, \frac{N_x+1}{2}$ and let the complex conjugate property define the values of $\tilde{\lambda}_p$ for $p = \frac{N_x+3}{2}, \ldots, N_x$. Since the eigenvalues have been designed around the NIMC scheme which uses the 1D DFT basis in (3.13) as eigenvectors and they possess the required complex conjugate property, the eigenvalues will be trialed with these eigenvectors. We will refer to this scheme as the *Modified NIMC* (MNIMC) scheme.

Definition 3.7 (The MNIMC scheme)**.** *Let Assumptions 3.2 hold true with* $\mathbf{U}^n$ *replaced by* $\tilde{\mathbf{U}}^n$ *to mark the different scheme. Define the matrix* $\tilde{M} \in \mathbb{R}^{N_x \times N_x}$ *where* N_x *is odd, by* $\tilde{M} := V\tilde{\Lambda}V^*$*, where* V *is defined in Section 3.3.1 and* $\tilde{\Lambda} := diag(\tilde{\lambda}_p)$ *the diagonal matrix of eigenvalues of the scheme,* $\tilde{\lambda}_p \in \mathbb{C}$ *for* $p = 1, \ldots, N_x$*. The eigenvalues of the scheme are defined by* $\tilde{\lambda}_p = e^{i\tilde{\theta}_p}$ *such that,*

$$\tilde{\theta}_p = \begin{cases} \frac{-2\pi i(p-1)sgn(\mu)h}{N_x}, & \text{for } p = 1, \ldots, \frac{N_x+1}{2}, \\ \frac{2\pi i(N_x-p+1)sgn(\mu)h}{N_x}, & \text{for } p = \frac{N_x+3}{2}, \ldots, N_x. \end{cases} \tag{3.37}$$

The scheme is implemented by multiplying $\tilde{U}^n$ *by the matrix* $\tilde{M}$ *to move the state of the system forward* Δt *in time, ie:* $\tilde{U}^{n+1} = \tilde{M}\tilde{U}^n$*.*

When $h = 1$, $\tilde{M} = M_{NIMC}$ and is also equal to the matrices implementing the Upwind, Preissman Box and Lax-Wendroff schemes when considered using $h = 1$. If we consider $h = \frac{q}{b}$, $q, b \in \mathbb{N}$ such that $\gcd(q, b) = 1$, we notice that,

$$\tilde{\lambda}_p^b = \lambda_p^q, \tag{3.38}$$

where λ_p is an eigenvalue of the NIMC scheme ($h_{NIMC} = 1$) for $p = 1, \ldots, N_x$. This gives us that by applying the MNIMC scheme b-times, we achieve the same result as applying the NIMC scheme q-times. We will now investigate the properties of the MNIMC scheme to understand its limitations for solving problem (3.1).

Consistency, convergence and numerical stability

In order to prove the consistency of the MNIMC scheme, rather than substituting the true solution into its schematic and then performing Taylor expansions, we will use an alternative method. This is because the schematic for the scheme makes use of every grid point in the domain (see Section 3.7.4), so the Taylor expansion method would make the process very complicated. Instead, we will use Fourier series to prove the consistency of the scheme in the following Lemma. This Lemma requires the use of Lemma 4.3 of Chapter 4 and the following definition.

Definition 3.8 (Regularity)**.** *A T-periodic function $T \in \mathbb{R}^+$, $f : \mathbb{R} \to \mathbb{R}$ such that $x \mapsto f(x)$ is defined as having regularity $r \in \mathbb{N}_0$, if r denotes the maximum number of times $f(x)$ can be differentiated such that $f^{(\alpha)}(x)$ is continuous and piecewise differentiable on $(0, T)$ for all $\alpha = 0, \ldots, r-1$ and $f^{(r)}(x)$ is piecewise continuous on $(0, T)$.*

The regularity of $u_0(x)$, provides an indicator for the smoothness of our initial condition and allows us to calculate the consistency of the MNIMC scheme.

Lemma 3.9. *Suppose the initial condition $u_0(x)$ for problem (3.1), has regularity $r \in \mathbb{N}_0$ over $(0, 1)$ and satisfies the conditions of Theorem 3.1 so it has a convergent Fourier series. Also let the conditions in Assumptions 3.2 hold true, so the MNIMC scheme can be defined as in Definition 3.7. Set the CFL number $h \in \mathbb{R}^+$ to be a fixed constant. Then the truncation error for the MNIMC scheme is such that,*

$$\tau_{j-1}^{n+1} = \mathcal{O}(\Delta x^r), \tag{3.39}$$

for all $n \in \mathbb{N}_0$ and $j = 1, \ldots, N_x$. Then for sufficiently smooth functions such that $r \in \mathbb{N}$,

$$\tau_{j-1}^{n+1} \to 0 \text{ and } \Delta t \to 0 \text{ as } \Delta x \to 0,$$

for all $n \in \mathbb{N}_0$ and $j = 1, \ldots, N_x$.

Proof. Applying one application of the MNIMC, results in $\tilde{\mathbf{U}}^{n+1} = \tilde{M}\tilde{\mathbf{U}}^n$ for fixed h.

Therefore,

$$\tilde{U}_{j-1}^{n+1} = \sum_{p=1}^{N_x} \left\{\tilde{M}\right\}_{j,p} \tilde{U}_{p-1}^{n},$$

for $j = 1, \ldots, N_x$, where $\left\{\tilde{M}\right\}_{j,p}$ denotes the (j,p)th element of the matrix $\tilde{M}$. If we now substitute in the true solution, we obtain,

$$\tau_{j-1}^{n+1} = u(x_{j-1}, t^{n+1}) - \sum_{p=1}^{N_x} \left\{\tilde{M}\right\}_{j,p} u(x_{p-1}, t^{n}), \tag{3.40}$$

for $j = 1, \ldots, N_x$.

Consider the Fourier series for $u(x,t)$ in (3.30) at time t^n. Then by direct calculation using the eigenvalue decomposition of $\tilde{M}$, or by Section 3.5 where c_k was found to be propagated by the $([k]_{N_x}+1)$th eigenvalue of the scheme, we derive,

$$\sum_{p=1}^{N_x} \left\{\tilde{M}\right\}_{j,p} u(x_{p-1}, t^{n}) = \sum_{k=-\infty}^{\infty} c_k \tilde{\lambda}_{[k]_{N_x}+1} e^{2\pi i k(x_{j-1} - \mu t^n)}, \tag{3.41}$$

for $j = 1, \ldots, N_x$ and $r \in \mathbb{N}$. This is because if $r \in \mathbb{N}$, $u_0(x)$ is continuous and hence equal to its Fourier series, resulting in the same to be true for $u(x,t)$. Then by (3.41) and (3.30) at time t^{n+1},

$$\tau_{j-1}^{n+1} = \sum_{k=-\infty}^{\infty} c_k e^{2\pi i k(x_{j-1} - \mu t^n)} \left(e^{-2\pi i k \mu \Delta t} - \lambda_{[k]_{N_x}+1} \right), \tag{3.42}$$

for $j = 1, \ldots, N_x$. Taking the absolute value of the truncation error,

$$\begin{aligned}
|\tau_{j-1}^{n+1}| &\leq \sum_{k=-\infty}^{\infty} |c_k| \left| e^{-2\pi i k \mu \Delta t} - \tilde{\lambda}_{[k]_{N_x}+1} \right|, \\
&= \sum_{p=1}^{N_x} \sum_{k=-\infty}^{\infty} |c_{p-1+kN_x}| \left| e^{-2\pi i (p-1+kN_x)\mu \Delta t} - \tilde{\lambda}_p \right|, \\
&\leq 2 \sum_{p=1}^{\frac{N_x+1}{2}} \sum_{k=1}^{\infty} (|c_{p-1+kN_x}| + |c_{p-1-kN_x}|) + \sum_{p=1}^{\frac{N_x+1}{2}} |c_{p-1}| \left| e^{-2\pi i (p-1)\mu \Delta t} - \tilde{\lambda}_p \right| \\
&\quad + 2 \sum_{p=\frac{N_x+3}{2}}^{N_x} \sum_{k=1}^{\infty} (|c_{p-1-N_x-kN_x}| + |c_{p-1-N_x+kN_x}|) \\
&\quad + \sum_{p=\frac{N_x+3}{2}}^{N_x} |c_{p-1-N_x}| \left| e^{2\pi i (N_x-p+1)\mu \Delta t} - \tilde{\lambda}_p \right|, \\
&\leq 2 \sum_{p=-\frac{N_x-3}{2}}^{\frac{N_x+1}{2}} \frac{D_3}{N_x^{r+1}}, \\
&\quad \text{by the proof of Lemma 4.3 and the definition of } \tilde{\lambda}_p, \\
&= \frac{2D_3}{N_x^r}, \qquad (3.43)
\end{aligned}$$

for $j = 1, \ldots, N_x$, where D_3 is a finite constant independent of N_x. As h is fixed, $\Delta t = \frac{h \Delta x}{\mu}$ is a function of Δx. As $\Delta x \to 0$, this results in $\Delta t \to 0$. Therefore, when the initial condition $u_0(x)$ is sufficiently smooth such that $r \in \mathbb{N}$,

$$|\tau_{j-1}^{n+1}| \leq 2D_3 \Delta x^r \to 0, \quad \text{as } \Delta x \to 0. \qquad (3.44)$$

□

The eigenvalues of the MNIMC scheme have unit magnitude for any value of h. This means that the scheme is always numerically stable. Then by the *Lax-Richtmyer Equivalence Theorem* [14, 71], the scheme is always convergent for a sufficiently smooth initial condition. Hence, the scheme can be used to solve problem (3.1) for any value of $h \in \mathbb{R}^+$. The exponential form of the eigenvalues also means that they are never zero, so the matrix implementing the scheme is always invertible.

3.7.4 Implementing the MNIMC scheme

In the last few Sections, we have shown that the scheme defined in Section 3.7.3, satisfies all the requirements we set out for it. However, by defining the scheme through its eigenvalues and eigenvectors, we do not have a schematic for the scheme as for the

Upwind, Preissman Box and Lax-Wendroff schemes. Possessing the schematic will allow us to assess the structure of the matrix $\tilde{M}$ implementing the scheme. Using the relationship $\tilde{\mathbf{U}}^{n+1} = \tilde{M}\tilde{\mathbf{U}}^n$ for all $n \in \mathbb{N}_0$, we can formulate the schematic equation for the scheme,

$$\tilde{U}_j^{n+1} = \frac{1}{N_x} \sum_{k=0}^{N_x-1} \left\{ 1 + 2 \sum_{p=2}^{\frac{N_x+1}{2}} \cos\left[\frac{2\pi(p-1)(j-k-\text{sgn}(\mu)h)}{N_x} \right] \right\} \tilde{U}_k^n, \tag{3.45}$$

for $j = 0, \ldots, N_x - 1$. This schematic constructs an explicit finite difference scheme implemented by the matrix $\tilde{M}$. The scheme uses the state of the system at every grid point in space at time t^n, to construct the state of the system at each grid point in space at time t^{n+1}. This gives the matrix $\tilde{M}$ a potentially non-zero value in every entry, unlike the matrices implementing the Upwind, Preissman Box and Lax-Wendroff schemes. The matrix $\tilde{M}$ is also a circulant matrix due to the circulant boundary conditions of the problem.

Figure 3.2 shows the results of applying the MNIMC scheme to the discrete sample of the 1D square function initial condition in (4.28), when $h = 0.5$. When t is an odd multiple of Δt, the scheme introduces an error into the numerical solution. This can be seen through the oscillations introduced into the numerical solution. When t is an even multiple of Δt, we can see that these errors are no longer present in the numerical solution. When implementing the scheme for $h = 1$, we find that the oscillations are not present. As the only errors introduced into the numerical solution by the scheme are those due to aliasing when $h \in \mathbb{R}^+ \setminus \mathbb{N}$, this must be the cause of the oscillations. However, there appears to be some periodic nature to this error. This is investigated in Lemma 3.12 of Section 3.9.

In the next Section, we consider using the Fourier series for the true solution to problem (3.1), to guide our choice in eigenvalues, for creating a scheme satisfying the conditions set out in Section 3.7.3. We can then compare the scheme this creates, with the MNIMC scheme.

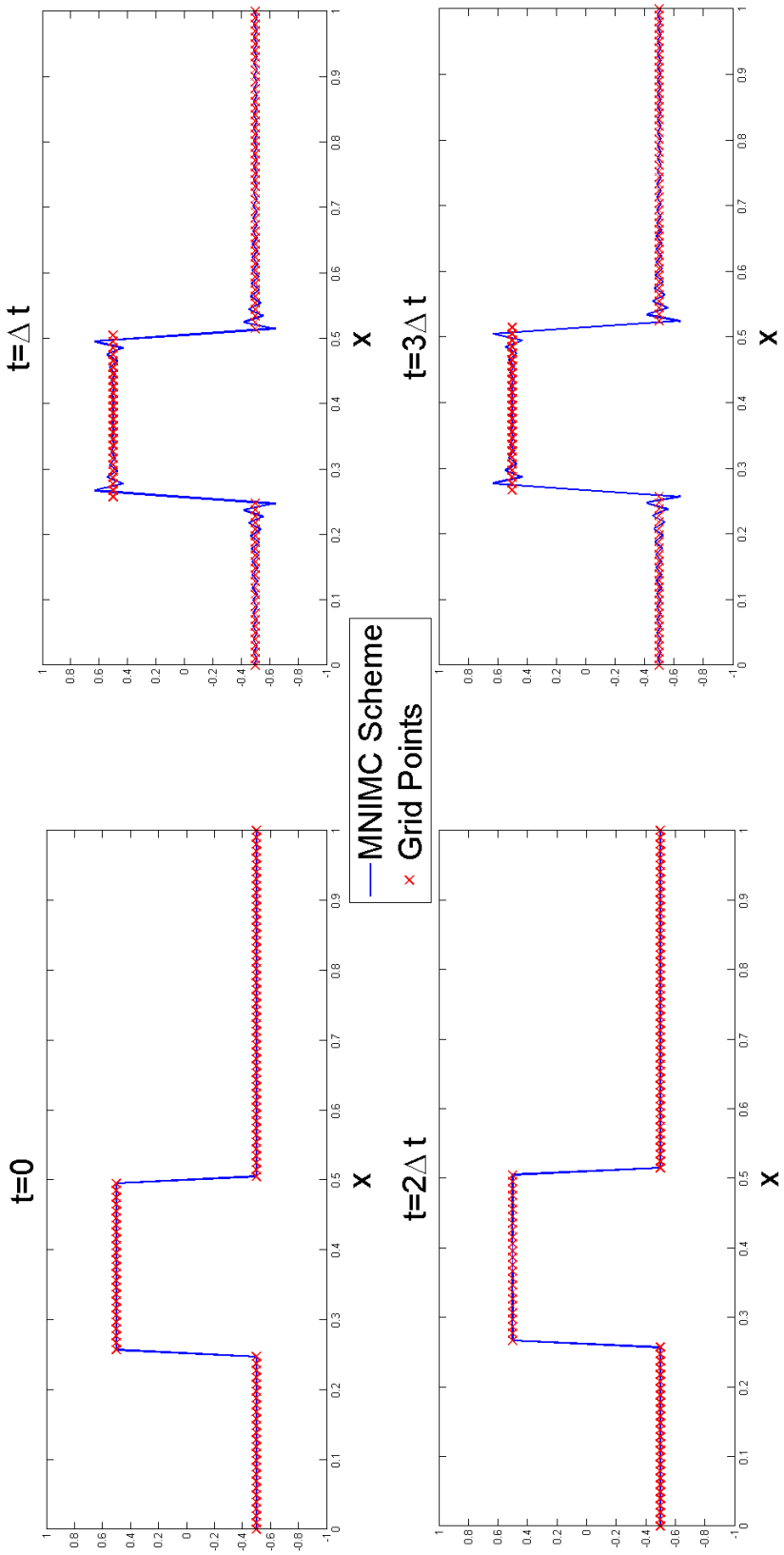


Figure 3.2: *The MNIMC finite difference scheme applied to the 1D square function initial condition in (4.28), for* $h = 0.5$, $\mu = 1$, $N_x = 101$ *and* $\Delta t = \frac{1}{202}$. *Here we can see the shifted* $2\Delta t$*-periodic nature of the aliasing error present in the scheme due to the denominator of* h *being equal to two.*

3.7.5 Defining a finite difference scheme using a Fourier series

In this Section, we will use the Fourier series for the analytical solution to problem (3.1), to define a scheme satisfying the properties set out in Section 3.7.3. Consider the Fourier series for $u(x,t)$ in (3.30). As discussed in Section 3.5, $g_k(1)$ multiplies the coefficient c_k of the Fourier series, to propagate the Fourier series forward Δt in time. In the case of the linear advection equation, $g_k(1) = e^{-2\pi ik\mu\Delta t}$. We require a scheme that propagates the resolvable wavenumber components of the Fourier series solution forward Δt in time, without introducing numerical dissipation or dispersion. As $g_k(1)$ is the correct coefficient to propagate the kth wavenumber component of the Fourier series, it seems a reasonable approach to choose the eigenvalues of the scheme to be $g_k(1)$ where k is the resolvable wavenumber component corresponding to that eigenvalue. We again remind the reader here that the negative resolvable wavenumber components of the Fourier series, correspond to eigenvalues λ_p for $p = \lfloor \frac{N_x}{2} \rfloor + 2, \ldots, N_x$. This can only be done when N_x is odd as the eigenvalue corresponding to $\mathbf{v}_{\frac{N_x}{2}+1}$ is required to be real, but the corresponding $g_{\frac{N_x}{2}}(1)$ is complex. This results in choosing N_x to be odd and

$$\lambda_p = \begin{cases} e^{-2\pi i(p-1)\mu\Delta t}, & \text{for } p = 1, \ldots, \frac{N_x+1}{2}, \\ e^{2\pi i(N_x-p+1)\mu\Delta t}, & \text{for } p = \frac{N_x+3}{2}, \ldots, N_x. \end{cases} \tag{3.46}$$

$$= \begin{cases} e^{\frac{-2\pi i(p-1)\mathrm{sgn}(\mu)h}{N_x}}, & \text{for } p = 1, \ldots, \frac{N_x+1}{2}, \\ e^{\frac{2\pi i(N_x-p+1)\mathrm{sgn}(\mu)h}{N_x}}, & \text{for } p = \frac{N_x+3}{2}, \ldots, N_x. \end{cases} \tag{3.47}$$

The CFL number gives that $|\mu|\Delta t = \frac{h}{N_x}$. As a result, the scheme we have defined here using the Fourier series for the analytical solution, has created the MNIMC scheme defined in Section 3.7.3. This demonstrates that the MNIMC scheme is a sensible scheme to define.

In the following Section, we will use the MNIMC scheme to aid us in identifying the numerically dissipative and dispersive properties of finite difference schemes, for solving problem (3.1). This will be achieved through the construction of metrics to determine the numerically dissipative and dispersive properties of schemes with respect to the resolvable wavenumber components of the solution.

3.8 Dissipative and dispersive metrics

When choosing a finite difference scheme to solve the 1D linear advection problem, it is important to understand the numerical model error associated with the scheme. The numerically dissipative and dispersive properties vary between each scheme and are dependent on the CFL number.

There is currently no satisfactory method available to assess the numerically dissipative and dispersive properties of a scheme, to judge whether it is appropriate for the

task. The damping factor and relative phase in Sections 3.5.2 and 3.5.3 respectively, provide a guide as to how the schemes attenuate or amplify and speed up or slow down wavenumber components of the numerical solution, respectively. These are calculated for each CFL number and produce values for each wavenumber. These ratios are useful once possible CFL numbers have been chosen that could yield schemes with the required properties. A method for choosing these CFL numbers is required.

To this end, the following definitions for a *dissipative metric* and a *dispersive metric* are defined. They provide a way to gauge the numerically dissipative and numerically dispersive properties of the resolvable wavenumbers of a scheme, in comparison to a reference scheme for comparison. The reference scheme needs to be the same for each considered scheme, for a fair comparison. If $g_k(1)$ is known, we advocate the use of

$$\lambda_p = \begin{cases} g_{p-1}(1), & \text{for } p = 1, \ldots, \frac{N_x+1}{2}, \\ g_{-N_x+p-1}(1), & \text{for } p = \frac{N_x+3}{2}, \ldots, N_x, \end{cases}$$

for a 1D problem, as these are the eigenvalues required to correctly propagate the resolvable wavenumber components of the numerical solution. This creates the MNIMC scheme as a reference scheme for comparison. However, any scheme could be used as a reference scheme for comparison if required.

The numerically dissipative and dispersive metrics need to provide a value that indicates, when two values are compared, that one scheme is more numerically dissipative or dispersive respectively than another. In order to allow this value to come about, the numerically dissipative and dispersive properties of each scheme need to be compared against the same property for the chosen reference scheme.

The metrics will be constructed using only the resolvable wavenumber components of the schemes. This is because the eigenvalues directly influence the propagation of these wavenumber components, whilst they influence unresolvable wavenumber components through aliasing. The numerically dissipative metric needs to compare the magnitude of the eigenvalue of a scheme against the eigenvalue of a chosen reference scheme for the same wavenumber component. Similarly, for the numerically dispersive metric, but through the comparison of the phases of the schemes. As a result, the considered scheme and the reference scheme must have the same value of N_x. The time step Δt must also be equal so the schemes move the solution forward in time equally. As the schemes are solving the same problem, the same characteristic speed is associated with each one, resulting in the same CFL number. As we are interested in how the metrics change with respect to the CFL number h, the eigenvalues are viewed as functions of h.

3.8.1 The dissipative metric

The numerically dissipative properties of a finite difference scheme used to solve problem (3.1), for a given CFL number, are found in the magnitude of the eigenvalues of

the scheme. The damping factor in Section 3.5.2 is one possibility for comparing the magnitude of the eigenvalues of a scheme against those of a reference scheme. However, a limitation of the damping factor is that when the magnitude of an eigenvalue of the reference scheme is zero, the damping factor is undefined. Instead we use the formulation of the *Frobenius norm* to define the numerically dissipative metric. The Frobenius norm is defined by $\|\cdot\|_F : \mathbb{C}^{m\times n} \to \mathbb{R}$ such that [65],

$$X \mapsto \|X\|_F := \left(\sum_{j=1}^{m} \sum_{k=1}^{n} |\{X\}_{j,k}|^2 \right)^{\frac{1}{2}}$$

Using this formulation, we construct the numerically dissipative metric from,

$$\|\Lambda\|_F^2 = \sum_{p=1}^{N_x} |\lambda_p|^2. \tag{3.48}$$

As the eigenvalues of the schemes have the complex conjugate property that $\overline{\lambda_p} = \lambda_{N_x-p+2}$ for $p = 2, \ldots, N_x$, only the eigenvalues for $p = 1, \ldots, \lfloor \frac{N_x}{2} \rfloor + 1$ are required. Calculating the difference between (3.48) for the considered scheme and (3.48) for the reference scheme, calculates the sum of the difference between the square of the magnitudes, for corresponding wavenumber components.

Definition 3.10 (Dissipative Metric)**.** *Define two finite difference schemes for solving problem (3.1), using the same spatial step size $\Delta x > 0$, and temporal step size $\Delta t > 0$. Let the eigenvalues of the scheme we wish to find the metric for, be denoted by $\left\{\lambda_p^{(1)}(h)\right\}_{p=1}^{N_x}$. Also let the eigenvalues of the reference scheme be denoted by $\left\{\lambda_p^{(2)}(h)\right\}_{p=1}^{N_x}$. Let the pth eigenvalue of each scheme correspond to the pth eigenvector of the 1D DFT basis, as defined in Section 3.3. Define the vectors $\mathbf{z}_1(h), \mathbf{z}_2(h) \in \mathbb{R}^{N_x}$ such that $[\mathbf{z}_j(h)]_p = |\lambda_p^{(j)}(h)|^2$ for $p = 1, \ldots, N_x$ and $j = 1, 2$. Then the numerically dissipative metric is defined by $d_{dissipative} : \mathbb{R}^{N_x} \times \mathbb{R}^{N_x} \to \mathbb{R}$, such that,*

$$\begin{aligned} d_{dissipative}(\mathbf{z}_1(h), \mathbf{z}_2(h)) &= \frac{1}{\lfloor \frac{N_x}{2} \rfloor + 1} \sum_{p=1}^{\lfloor \frac{N_x}{2} \rfloor + 1} \left| [\mathbf{z}_1(h)]_p - [\mathbf{z}_2(h)]_p \right|, \\ &= \frac{1}{\lfloor \frac{N_x}{2} \rfloor + 1} \sum_{p=1}^{\lfloor \frac{N_x}{2} \rfloor + 1} \left| |\lambda_p^{(1)}(h)|^2 - |\lambda_p^{(2)}(h)|^2 \right|. \end{aligned} \tag{3.49}$$

The dissipative metric is normalised with respect to the number of wavenumber components. The metric is always greater than or equal to zero as the sum is con-

structed from positive values by design. When the scheme we wish to find the numerically dissipative metric for has eigenvalues such that $|\lambda_p^{(1)}| = |\lambda_p^{(2)}|$ for all $p = 1, \ldots, N_x$, the numerically dissipative metric is zero. We also find that when the reference scheme is numerically non-dissipative, the higher the value of the numerically dissipative metric, the greater the numerical dissipation of the scheme being tested. Definition 3.10 defined a metric on $\mathbb{R}$.

In the case of the 1D linear advection problem, the numerically non-dissipative MNIMC scheme is chosen as the reference scheme for comparison, requiring N_x to be odd. The eigenvalues of the MNIMC scheme all have magnitude one so,

$$d_{dissipative}(h) = \frac{2}{N_x + 1} \sum_{p=1}^{\frac{N_x+1}{2}} \left| |\lambda_p(h)|^2 - 1 \right|. \tag{3.50}$$

The numerically dissipative metric in (3.50) for the Upwind, Preissman Box and Lax-Wendroff schemes is shown in Figure 3.3, for various values of the CFL number, such that the schemes are all numerically stable ie: $0 < h \leq 1$. The metric is also shown for the MNIMC for comparison. As the schemes are numerically stable, the magnitude of their eigenvalues is less than or equal to one, resulting in the dissipative metric being bounded by one.

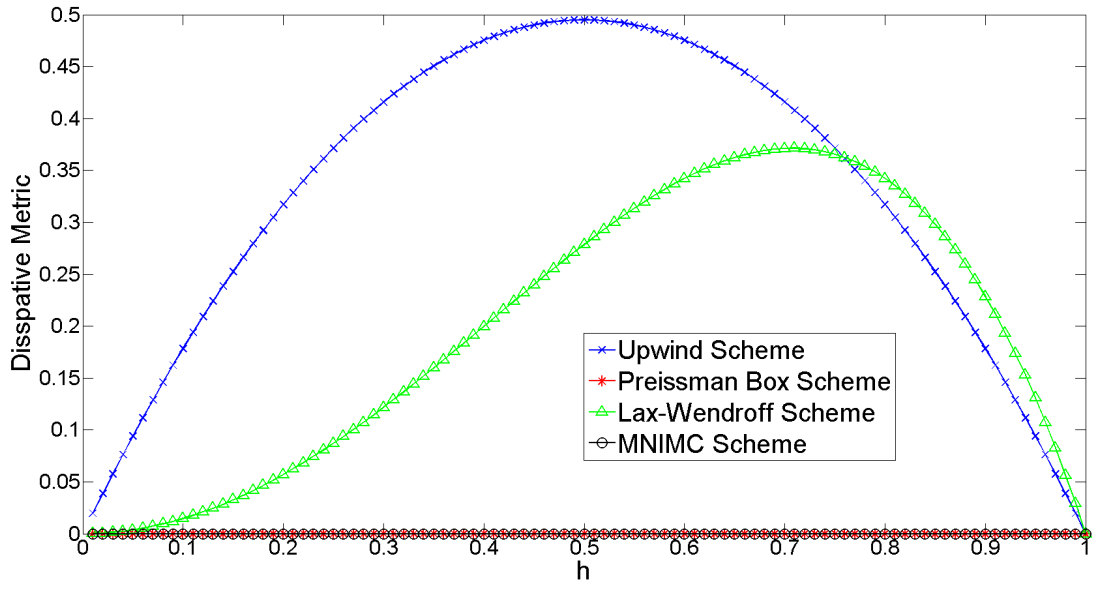


Figure 3.3: *The dissipative metric in (3.50) for the Upwind, Preissman Box, Lax-Wendroff and MNIMC schemes when* $N_x = 101$ *and* $\mu = 1$. *The CFL number is considered for* $0 < h \leq 1$.

Examining Figure 3.3, we see that the dissipative metric in (3.50) is zero for all considered CFL numbers, for both the MNIMC and Preissman Box schemes. This shows that both schemes are numerically non-dissipative with respect to the resolvable wavenumber components of the numerical solution, as already identified for these schemes in Table 3.2. This is the case for the MNIMC scheme by definition. The

Upwind and Lax-Wendroff schemes have values that are greater than zero for nearly all considered values of the CFL number. The dissipative metric for both of these schemes is zero when $h = 1$, showing these schemes are numerically non-dissipative at this value. The Upwind scheme appears to have numerically dissipative properties which are symmetric about $h = 0.5$. The metric for the Lax-Wendroff scheme is skewed to the right of $h = 0.5$, showing that choosing a CFL number around 0.75 will result in more numerical dissipation in the scheme.

3.8.2 The dispersive metric

The numerically dispersive properties of a finite difference scheme used to solve problem (3.1), for a given CFL number, are found in the phase of the eigenvalues of the scheme. Just as for the numerically dissipative metric, the relative phase in Section 3.5.3 is one possibility for comparing the phases of the eigenvalues of a scheme, against those of a reference scheme. However, we also have the limitation that when the phase of an eigenvalue of the reference scheme is zero, the relative phase cannot be defined. Instead we choose to structure the metric similarly to the numerically dissipative metric, favouring the sum of differences of phases.

As before, by the complex conjugate property of the eigenvalues, only the eigenvalues for $p = 1, \ldots, \lfloor \frac{N_x}{2} \rfloor + 1$ will form a part of the metric. There is the additional problem that given any phase θ_p, $e^{i\theta_p} = e^{i(\theta_p + 2\pi k)}$ where $k \in \mathbb{Z}$. It is possible for the phase θ_p to be the incorrect phase, but still lead to the correct phase change for the corresponding wavenumber component. However, when determining the phase of the Upwind, Preissman Box and Lax-Wendroff schemes, from their eigenvalues, $\theta_p(h) = \arctan\left[\frac{\mathrm{Im}(\lambda_p(h))}{\mathrm{Re}(\lambda_p(h))}\right]$, so $\theta_p \in [-\pi, \pi)$. To make the comparison fair, we require that the phases for the scheme the metric is being found for and the reference scheme, both be mapped to the domain $[-\pi, \pi)$.

Definition 3.11 (Dispersive Metric)**.** *Define two finite difference schemes for solving problem (3.1), using the same spatial step size $\Delta x > 0$, and temporal step size $\Delta t > 0$. Let the eigenvalues of the scheme we wish to find the metric for, be denoted by $\left\{\lambda_p^{(1)}(h)\right\}_{p=1}^{N_x}$. Also let the eigenvalues of the reference scheme be denoted by $\left\{\lambda_p^{(2)}(h)\right\}_{p=1}^{N_x}$. Let the pth eigenvalue of each scheme correspond to the pth eigenvector of the 1D DFT basis, as defined in Section 3.3. Define the vectors $\mathbf{z}_1(h), \mathbf{z}_2(h) \in \mathbb{R}^{N_x}$ such that $[\mathbf{z}_j(h)]_p = \theta_p^{(j)}(h)$, where $\theta_p^{(j)}(h) \in [-\pi, \pi)$ is the phase of the eigenvalue $\lambda_p^{(j)}(h)$ for $p = 1, \ldots, N_x$ and $j = 1, 2$. Then the numerically dispersive metric is*

defined by $d_{dispersive} : \mathbb{R}^{N_x} \times \mathbb{R}^{N_x} \to \mathbb{R}$, *such that,*

$$\begin{aligned} d_{dispersive}(\mathbf{z}_1(h), \mathbf{z}_2(h)) &= \frac{1}{2\pi \left\lfloor \frac{N_x}{2} \right\rfloor + 1} \sum_{p=1}^{\left\lfloor \frac{N_x}{2} \right\rfloor + 1} \left| [\mathbf{z}_1(h)]_p - [\mathbf{z}_2(h)]_p \right|, \\ &= \frac{1}{2\pi \left\lfloor \frac{N_x}{2} \right\rfloor + 1} \sum_{p=1}^{\left\lfloor \frac{N_x}{2} \right\rfloor + 1} \left| \theta_p^{(1)}(h) - \theta_p^{(2)}(h) \right|. \end{aligned} \tag{3.51}$$

The metric is normalised with respect to the number of wavenumber components and the bound on the difference between the phases ie: 2π. The dispersive metric is always greater than or equal to zero as the sum is constructed from positive values by design. When the scheme we wish to find the numerically dispersive metric for has eigenvalues such that $\theta_p^{(1)} = \theta_p^{(2)}$ for all $p = 1, \ldots, N_x$, the numerically dispersive metric is zero. We also find that when the reference scheme is numerically non-dispersive with respect to the resolvable wavenumber components, the higher the value of the numerically dispersive metric, the greater the numerical dispersion of the scheme being tested. Definition 3.11 is a metric on $\mathbb{R}$.

In the case of the 1D linear advection problem, the numerically non-dispersive MNIMC scheme is chosen as the reference scheme for comparison,

$$d_{dispersive}(h) = \frac{2}{\pi(N_x + 1)} \sum_{p=1}^{\frac{N_x+1}{2}} \left| \theta_p(h) - \left(\frac{-2\pi i(p-1)h}{N_x} \right) \right|. \tag{3.52}$$

The numerically dispersive metric in (3.52) for the Upwind, Preissman Box and Lax-Wendroff schemes is shown in Figure 3.4, for various values of the CFL number, such that the schemes are all numerically stable ie: $0 < h \leq 1$. The metric is also shown for the MNIMC for comparison.

The metric for the MNIMC scheme shows that the scheme is always numerically non-dispersive with respect to the resolvable wavenumber components, as expected by its definition. Examining the metric for the Upwind scheme, we see the same symmetry about $h = 0.5$ as seen in the numerically dissipative metric. It also shows that as already discovered, the Upwind scheme is numerically non-dispersive with respect to the resolvable wavenumber components for $h = 0.5$ and $h = 1$.

The numerically dispersive metric for the Preissman Box and Lax-Wendroff schemes are skewed slightly in alternate directions, with both schemes introducing numerical dispersion, apart from when $h = 1$. When considering a small CFL number ($h \ll 1$), the Upwind and Lax-Wendroff schemes introduce about the same amounts of numerical dispersion. For CFL numbers close to $h = 0.5$, choosing the Upwind scheme would limit the effects of numerical dispersion. When considering large CFL numbers ($h \approx 1$), the Lax-Wendroff scheme would introduce the least numerical dispersion into the numerical

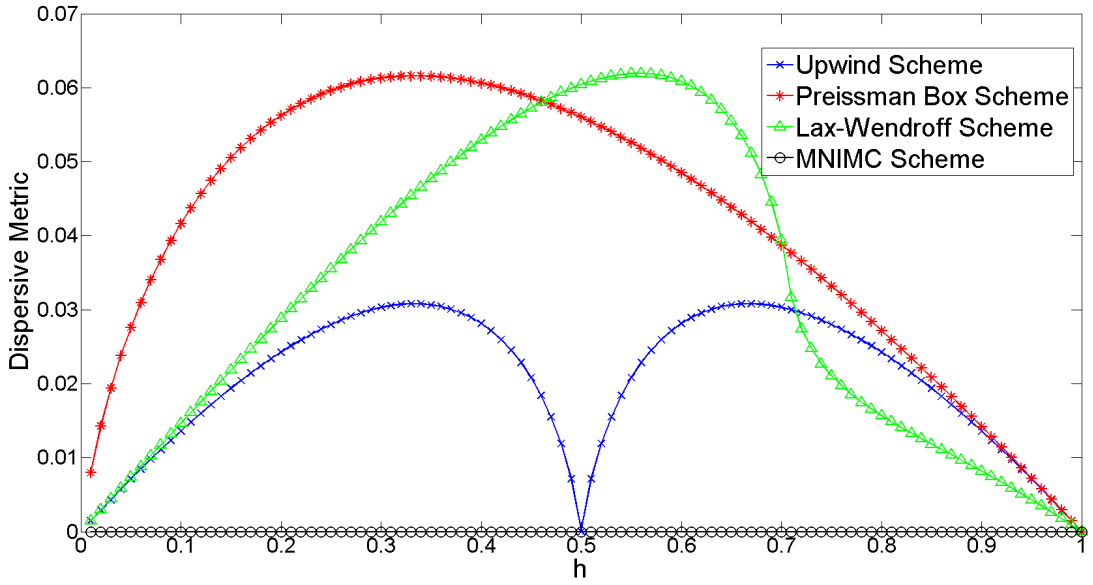


Figure 3.4: *The dispersive metric in (3.52) for the Upwind, Preissman Box, Lax-Wendroff and MNIMC schemes when* $N_x = 101$ *and* $\mu = 1$*. The CFL number is considered for* $0 < h \leq 1$*.*

solution.

The numerically dissipative and dispersive metrics can be used in combination to choose the best scheme for the considered problem. Once candidate CFL numbers have been chosen, the damping factor and relative phase can be used to learn more about the effects of the schemes on individual wavenumber components.

3.9 Aliasing errors in the MNIMC scheme

In Section 3.7.3 we defined the MNIMC scheme and identified that it is numerically non-dissipative with respect to all wavenumber components and numerically non-dispersive with respect to all resolvable wavenumber components, for any value of $h \in \mathbb{R}^+$. When $h \in \mathbb{R}^+ \setminus \mathbb{N}$ the scheme is numerically dispersive with respect to the unresolvable wavenumber components of the numerical solution. Whilst implementing the scheme in Section 3.7.4, we found that the scheme introduced anomalous oscillations into the numerical solution at periodic intervals in time, when $h \in \mathbb{R}^+ \setminus \mathbb{N}$. As the only error in the scheme is due to numerical dispersion in the unresolvable wavenumber components, it is this error that must be causing these oscillations.

Let $\tilde{\mathbf{x}}_0 \in \mathbb{R}^{N_x}$ denote the true initial condition $u_0(x)$, sampled at the spatial grid points $x_0, \ldots, x_{N_x-1}$ defined in Assumptions 3.2, such that $\{\tilde{\mathbf{x}}_0\}_j := u_0(x_{j-1})$. Now define $\tilde{\mathbf{x}}_l \in \mathbb{R}^{N_x}$ by $\tilde{\mathbf{x}}_l := \tilde{M}^l \tilde{\mathbf{x}}_0$ for all $l \in \mathbb{N}$, the lth state of the system generated by the MNIMC scheme. Then the *global error* in the MNIMC scheme is defined by,

$$\mathbf{r}_l := \tilde{\mathbf{y}}_l - \tilde{\mathbf{x}}_l = \tilde{\mathbf{y}}_l - \tilde{M}^l \tilde{\mathbf{x}}_0, \tag{3.53}$$

where $\mathbf{y}_l := \tilde{\mathbf{y}}_l$ denotes a perfect observation as defined in Section 2.3. As only aliasing errors are introduced by the MNIMC scheme, $\mathbf{r}_l$ can be viewed as an additive correction term to correct for aliasing errors in $\tilde{M}^l\tilde{\mathbf{x}}_0$ such that,

$$\tilde{\mathbf{y}}_l = \tilde{\mathbf{x}}_l + \mathbf{r}_l = \tilde{M}^l\tilde{\mathbf{x}}_0 + \mathbf{r}_l. \tag{3.54}$$

An additive correction term was also chosen by Daley [1] in his analysis of model errors. Choosing $h = 1$ results in $\tilde{M} = M_{NIMC}$ and consequently $\mathbf{r}_l = \mathbf{0}$ for all l. Lemma 3.12 provides insight into the properties of the aliasing error introduced by the MNIMC scheme.

Lemma 3.12. *Let the conditions in Assumptions 3.2 hold true so the MNIMC scheme can be defined as in Definition 3.7. Also, let $u_0(x)$ be bounded and piecewise continuous on $[0,1)$ and suppose the left- and right-hand derivatives of $u_0(x)$ exist for all $x \in [0,1)$.*

Additionally, consider the CFL number to be a rational number $h \in \mathbb{Q}^+$ expressed as $h = \frac{q}{b}$, $q, b \in \mathbb{N}$ such that $gcd(q,b) = 1$. Then the global error in the MNIMC scheme at time $l\Delta t$, defined by Equation (3.53), is such that,

$$\mathbf{r}_l = \begin{cases} \mathbf{0}, & \text{for } [l]_b = 0, \\ \tilde{M}^{l-[l]_b}\mathbf{r}_{[l]_b}, & \text{for } [l]_b = 1, \ldots, b-1, \end{cases} \tag{3.55}$$

for all $l \in \mathbb{N}_0$, where $[\cdot]_b$ denotes modulo b.

Proof. As we are investigating the MNIMC scheme, N_x must be odd. Rearranging (3.54) and applying the 1D DFT results in,

$$\mathcal{F}_p(\mathbf{r}_l) = \mathcal{F}_p(\tilde{\mathbf{y}}_l) - \tilde{\lambda}_p^l\mathcal{F}_p(\tilde{\mathbf{x}}_0) \tag{3.56}$$

for all $l \in \mathbb{N}_0$. The vector $\tilde{\mathbf{y}}_l$ contains a discrete sample of the true physical system sampled at each grid point in space, at time $l\Delta t$. Therefore,

$$[\tilde{\mathbf{y}}_l]_p = u(x_{p-1} - \mu l\Delta t, 0), \quad \text{for all } p = 1, \ldots, N_x. \tag{3.57}$$

We would like to use the Fourier series for $u_0(x)$ in (3.6), to represent $u(x - \mu l\Delta t, 0)$. Under the conditions of the Lemma, this Fourier series is convergent.

The function $u(x,0)$ is a periodic extension of the function $u_0(x)$. If $u_0(x)$ is continuous over $[0,1)$ and $\lim_{x\to 0^+} u_0(x) = \lim_{x\to 1^-} u_0(x)$, then under the conditions of the Lemma, the Fourier series for $u_0(x)$ is equal to the function $u(x,0)$ for all $x \in \mathbb{R}$. Then the Fourier series of $u_0(x)$ can be used to represent $u(x_{p-1} - \mu l\Delta t, 0)$.

However, if $\lim_{x\to 0^+} u_0(x) \neq \lim_{x\to 1^-} u_0(x)$ or $u_0(x)$ is piecewise continuous over $[0,1)$ then there exists a finite number of discontinuities in $u_0(x)$ which are periodically repeated in $u(x,0)$. These are jump discontinuities, by the conditions of the Lemma.

In this case the Fourier series for $u_0(x)$ converges to $u(x,0)$ at all points in the domain, except those where the function is discontinuous, where it converges to the midpoint of the jump discontinuity. In this instance, when none of the sample points are points of discontinuity, the Fourier series of $u_0(x)$ can be used to represent $u(x_{p-1}-\mu l\Delta t,0)$ as for the previous case. However when for a given l, the pth sample point coincides with a discontinuity, the Fourier series converges to the midpoint of the discontinuity and not $u(x_{p-1}-\mu l\Delta t,0)$. Then the Fourier series of $u_0(x)$ is not equal to $u(x_{p-1}-\mu l\Delta t,0)$, so cannot be used to represent this sample point.

In this instance, we define a new one-periodic function for each l, whose Fourier series can be used to represent the discontinuous function at each sample point. This function must be both continuous and have the same value as $u(x-\mu l\Delta t,0)$, at every sample point in space. This function is defined over $\left[-\mu l\Delta t-\frac{\Delta x}{2},1-\mu l\Delta t-\frac{\Delta x}{2}\right)$ by placing triangular functions into $u(x-\mu l\Delta t,0)$, with the apex coinciding with a point of discontinuity and either side extending $\frac{\Delta x}{2}$ back to the function $u(x-\mu l\Delta t,0)$. An example can be seen in Figure 3.5.

Define,

$$\hat{X}=\{x\in[0,1)|u(x,0)\text{ is a jump discontinuity}\}, \tag{3.58}$$

and $\hat{X}_l\subseteq\hat{X}$, for each $l\in\mathbb{N}_0$,

$$\hat{X}_l=\left\{\hat{x}\in\hat{X}|\exists p\in\{1,\dots,N_x\}\text{ such that }\hat{x}=[x_{p-1}-\mu l\Delta t]_1\right\}. \tag{3.59}$$

This set identifies the sample points within $[0,1)$ where discontinuities lie within $u(x-\mu l\Delta t,0)$. If there exists l such that $u(x_{p-1}-\mu l\Delta t,0)$ is a continuous point for all $p=1,\dots,N_x$, then $\hat{X}_l=\emptyset$. When $u(x,0)$ is a continuous function over $\mathbb{R}\times\{0\}$, $\hat{X}=\emptyset$, hence $\hat{X}_l=\emptyset$ for all $l\in\mathbb{N}_0$.

Consider $\hat{x}\in\hat{X}_l$, then there exists $p\in\{1,\dots,N_x\}$ such that,

$$\begin{aligned}
\hat{x} &= [x_{p-1}-\mu l\Delta t]_1,\\
&= [x_{p-1-\text{sgn}(\mu)h(l-[l]_b)}-\mu[l]_b\Delta t]_1,\ \text{ as } h=\frac{|\mu|\Delta t}{\Delta x},\\
&= [\alpha+x_{[p-1-\text{sgn}(\mu)h(l-[l]_b)]_{N_x}}-\mu[l]_b\Delta t]_1,\\
&\quad \text{for some } \alpha\in\mathbb{Z} \text{ such that}\\
&\quad p-1-\text{sgn}(\mu)h(l-[l]_b)-[p-1-h(l-[l]_b)]_{N_x}=\alpha N_x,\\
&= [x_{[p-1-\text{sgn}(\mu)h(l-[l]_b)]_{N_x}}-\mu[l]_b\Delta t]_1.
\end{aligned} \tag{3.60}$$

As $[p-1-\text{sgn}(\mu)h(l-[l]_b)]_{N_x}\in\{0,\dots,N_x-1\}$, there exists some $q\in\{1,\dots,N_x\}$ such that $q-1=[p-1-\text{sgn}(\mu)h(l-[l]_b)]_{N_x}$, hence $\hat{x}=[x_{q-1}-\mu[l]_b\Delta t]_1$. As a result, $\hat{x}\in\hat{X}_l\Leftrightarrow\hat{x}\in\hat{X}_{[l]_b}$, so $\hat{X}_l=\hat{X}_{[l]_b}$ for all $l\in\mathbb{N}_0$. As $\hat{X}_l=\hat{X}_{[l]_b}$, we have shown there are at most b different subsets of points in $u(x,0)$ over $[0,1)$, where a discontinuity is sampled, over time.

We will now use these subsets to define our new functions, $v_l:\mathbb{R}\to\mathbb{R}$ such that

$x \mapsto v_l(x)$ for all $l \in \mathbb{N}_0$ such that $\hat{X}_l \neq \emptyset$, using the points in $\hat{X}_l = \{\hat{x}_j\}_{j=1}^{|\hat{X}_l|}$,

$$v_l(x) = \begin{cases} \left[\frac{u(\hat{x}_j,0)-u(\hat{x}_j-\frac{\Delta x}{2},0)}{\frac{\Delta x}{2}}\right](x-\hat{x}_j+\frac{\Delta x}{2})+u\left(\hat{x}_j-\frac{\Delta x}{2},0\right), \\ \text{for } x \in \left[\hat{x}_j-\frac{\Delta x}{2},\hat{x}_j\right), \text{ where } \hat{x}_j \in \hat{X}_l, \\ \\ \left[\frac{u\left(\hat{x}_j+\frac{\Delta x}{2},0\right)-u(\hat{x}_j,0)}{\frac{\Delta x}{2}}\right](x-\hat{x}_j)+u(\hat{x}_j,0), \\ \text{for } x \in \left[\hat{x}_j,\hat{x}_j+\frac{\Delta x}{2}\right), \text{ where } \hat{x}_j \in \hat{X}_l, \\ \\ u(x,0), \\ \text{for } x \in \left[-\mu l\Delta t-\frac{\Delta x}{2},1-\mu l\Delta t-\frac{\Delta x}{2}\right) \setminus \bigcup_{j=1}^{|\hat{X}_l|}\left[\hat{x}_j-\frac{\Delta x}{2},\hat{x}_j+\frac{\Delta x}{2}\right), \end{cases} \tag{3.61}$$

and $v_l(x+1) = v_l(x)$ for all $x \in \mathbb{R}$. This creates a one-periodic function that is equal to $u(x,0)$, except at the points within a radius of $\frac{\Delta x}{2}$, from the points in $\hat{X}_l$. A linear interpolation is created over the discontinuous point $\hat{x}_j \in \hat{X}_l$, from $u(\hat{x}_j - \frac{\Delta x}{2}, 0)$ to $u(\hat{x}_j, 0)$ and from $u(\hat{x}_j, 0)$ to $u(\hat{x}_j + \frac{\Delta x}{2}, 0)$. This ensures there are no discontinuous sample points in $v_l(x)$ and $v_l([x_{p-1}-\mu l\Delta t]_1) = u([x_{p-1}-\mu l\Delta t]_1, 0)$ for all $p = 1, \ldots, N_x$. As $\hat{X}_l = \hat{X}_{[l]_b}$, $v_l(x) = v_{[l]_b}(x)$ for all l so there are at most b different functions $v_l(x)$.

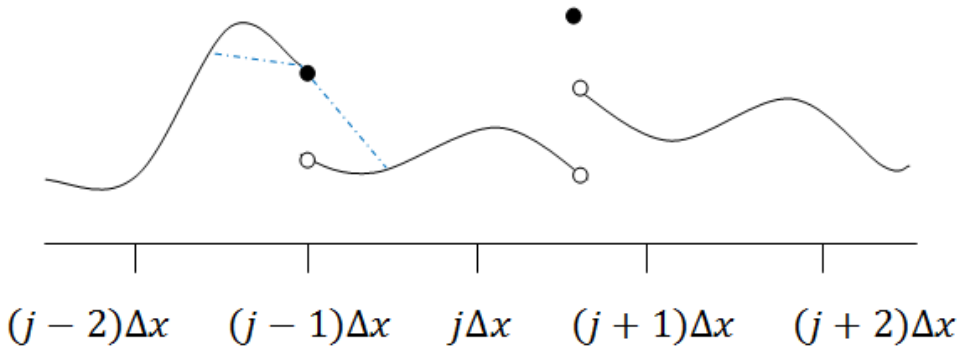


Figure 3.5: *The function plotted with a solid black line is a particular $u_0(x)$ for $x \in [(j-2)\Delta x, (j+2)\Delta x] \in [0,1)$, for some $j \in \mathbb{N}_0$, $2 \leq j \leq N_x - 3$. The function $v_0(x)$ is plotted over the same domain and is given by the function $u_0(x)$ except over $\left[(j-1)\Delta x - \frac{\Delta x}{2}, (j-1)\Delta x + \frac{\Delta x}{2}\right)$, where the function is defined by the broken blue line. The broken blue line represents a triangular function placed into the function $u_0(x)$ over the discontinuity at $(j-1)\Delta x$.*

Now define a Fourier series for $v_l(x)$ for $l = 0, \ldots, b-1$, such that $\hat{X}_l \neq \emptyset$,

$$v_l(x) = \sum_{k=-\infty}^{\infty} d_k^{(l)} e^{2\pi i k x}, \tag{3.62}$$

where $d_k^{(l)} \in \mathbb{C}$ for all $k \in \mathbb{Z}$. This is a convergent Fourier series for $v_l(x)$ by the conditions of the Lemma.

Next consider the 1D DFT applied to $\tilde{\mathbf{y}}_l$. When $\hat{X}_l = \emptyset$, the Fourier series of $u_0(x)$

should be considered as can be seen in the calculations below. The same calculations are carried out using the Fourier series for $v_l(x)$ when $\hat{X}_l \neq \emptyset$. We are able to do this as by showing that $v_l(x) = v_{[l]_b}(x)$, we have shown that $d_k^{(l)} = d_k^{([l]_b)}$ for all $k \in \mathbb{Z}$.

$$\begin{aligned}
\mathcal{F}_p(\tilde{\mathbf{y}}_l) &= \sqrt{N_x} \sum_{k=-\infty}^{\infty} c_{p-1+kN_x} e^{-2\pi i(p-1+kN_x)\mu l \Delta t}, \text{ by the Poisson summation,} \\
&= \sqrt{N_x} e^{\frac{-2\pi i(p-1)\text{sgn}(\mu)h(l-[l]_b)}{N_x}} \sum_{k=-\infty}^{\infty} c_{p-1+kN_x} e^{\frac{-2\pi i(p-1+kN_x)\mu[l]_b\Delta t}{N_x}}, \text{ as } h = \frac{|\mu|\Delta t}{\Delta x}, \\
&= e^{\frac{-2\pi i(p-1)\text{sgn}(\mu)h(l-[l]_b)}{N_x}} \mathcal{F}_p(\tilde{\mathbf{y}}_{[l]_b}). \qquad (3.63)
\end{aligned}$$

We note that as,

$$e^{\frac{-2\pi i(p-1)\text{sgn}(\mu)h(l-[l]_b)}{N_x}} = e^{\frac{-2\pi i(p-1+sN_x)\text{sgn}(\mu)h(l-[l]_b)}{N_x}},$$

for any $s \in \mathbb{Z}$, we can re-write (3.63) as follows,

$$\begin{aligned}
\mathcal{F}_p(\tilde{\mathbf{y}}_l) &= \begin{cases} e^{\frac{-2\pi i(p-1)\text{sgn}(\mu)h(l-[l]_b)}{N_x}} \mathcal{F}_p(\tilde{\mathbf{y}}_{[l]_b}), & \text{for } p = 1, \ldots, \frac{N_x+1}{2}, \\ e^{\frac{2\pi i(N_x-p+1)\text{sgn}(\mu)h(l-[l]_b)}{N_x}} \mathcal{F}_p(\tilde{\mathbf{y}}_{[l]_b}), & \text{for } p = \frac{N_x+3}{2}, \ldots, N_x, \end{cases} \\
&= \tilde{\lambda}_p^{\,l-[l]_b} \mathcal{F}_p(\tilde{\mathbf{y}}_{[l]_b}), \text{ for } p = 1, \ldots, N_x. \qquad (3.64)
\end{aligned}$$

We also have that,

$$\tilde{\lambda}_p^l \mathcal{F}_p(\tilde{\mathbf{x}}_0) = \tilde{\lambda}_p^{l-[l]_b} \tilde{\lambda}_p^{[l]_b} \mathcal{F}_p(\tilde{\mathbf{x}}_0), \text{ for } p = 1, \ldots, N_x. \qquad (3.65)$$

Therefore, substituting (3.64) and (3.65) into (3.56),

$$\begin{aligned}
& \mathcal{F}_p(\mathbf{r}_l) = \tilde{\lambda}_p^{l-[l]_b} \mathcal{F}_p(\mathbf{r}_{[l]_b}), \text{ for } p = 1, \ldots, N_x, \\
\Rightarrow\ & V^* \mathbf{r}_l = \tilde{\Lambda}^{l-[l]_b} V^* \mathbf{r}_{[l]_b}, \\
\Rightarrow\ & \mathbf{r}_l = \tilde{M}^{l-[l]_b} \mathbf{r}_{[l]_b}, \qquad (3.66)
\end{aligned}$$

for $l \in \mathbb{N}_0$. As $\tilde{\mathbf{y}}_0 = \tilde{\mathbf{x}}_0$, $\mathbf{r}_0 = \mathbf{0}$. Then by (3.66), when $[l]_b = 0$, $\mathbf{r}_l = \mathbf{0}$. Hence the result in (3.55).

□

In order to understand the result of this Lemma in Equation (3.55), we need to remember the result of Equation (3.38). This determined that by applying $\tilde{M}$ b-times, this was the same as applying the NIMC scheme q-times ($h_{NIMC} = 1$), where $h = \frac{q}{b}$, $b, q \in \mathbb{N}$ and $\gcd(b, q) = 1$ for the MNIMC scheme.

Raising the matrix $\tilde{M}$ to the power $l - [l]_b$ results in $\tilde{M}$ being raised to a power which

is an integer multiple of b. Suppose $l-[l]_b = sb$ for some $s \in \mathbb{N}_0$, then $\tilde{M}^{l-[l]_b} = M_{NIMC}^{sq}$. Applying this matrix to $\mathbf{r}_{[l]_b}$ shifts it $sq\Delta x$ in space, preserving its shape. Then $\mathbf{r}_l$ is shifted an integer number of discretisation points in space. This means that the observation points on the domain $[0, 1)$, due to the one-periodic nature of $u(x, t)$, sample the same points of the function as they did when the function was last moved an integer multiple of Δx in space. This gives the error in the MNIMC scheme a shifted $b\Delta t$-periodic nature. This is what was seen in Figure 3.2, where $b = 2$.

The impact of aliasing errors on the results of the MNIMC scheme, means that the scheme is unable to provide perfect observations every Δt in time, as $\mathbf{r}_l$ is unknown for $l = 1, \ldots, b-1$. However using the MNIMC scheme to construct our perfect observations, as in Equation (3.54), makes our algebraic analysis easier. As the *circshift* function of MATLAB®[74] allows perfect observations to be generated numerically every Δx in space as described in Section 3.7.2, we will use this method to generate perfect observations numerically.

Consider the MNIMC scheme and swap the eigenvalues corresponding to a conjugate pair of the 1D DFT basis, eg: swap λ_p and λ_{N_x-p+2} for some $p = 2, \ldots, N_x$, so that $(\lambda_{N_x-p+2}, \mathbf{v}_p)$ and $(\lambda_p, \mathbf{v}_{N_x-p+2})$ form eigenpairs of the scheme. Then this scheme possesses all the properties of the MNIMC scheme except it is not numerically non-dispersive with respect to the pth resolvable real wavenumber component. However, the shifted $b\Delta t$-periodic nature identified in Lemma 3.12 for the MNIMC scheme, still holds. This can be seen in Figure 3.6 where λ_4 and λ_{N_x-2} were swapped.

This is due to the form of the eigenvalues. Any number of these pairs of eigenvalues can be swapped to create a scheme that has a shifted $b\Delta t$-periodic nature, similar to that described by Lemma 3.12 for the MNIMC scheme. As a result, it is possible to define $2^{\frac{N_x-1}{2}}$ different schemes for solving problem (3.1), each possessing the shifted $b\Delta t$-periodic nature described by Lemma 3.12. However, only the definition of the MNIMC scheme produces a scheme that is numerically non-dispersive with respect to all resolvable wavenumber components of the numerical solution.

Now we have a way to construct our perfect observations algebraically and a way to generate them numerically using the *circshift* function of MATLAB®[74], we can consider our data assimilation problem. The following sections examine the impact of numerical model error from finite difference schemes, on the construction of the analysis vector, when performing strong constraint 4D-Var data assimilation for our 1D linear advection problem.

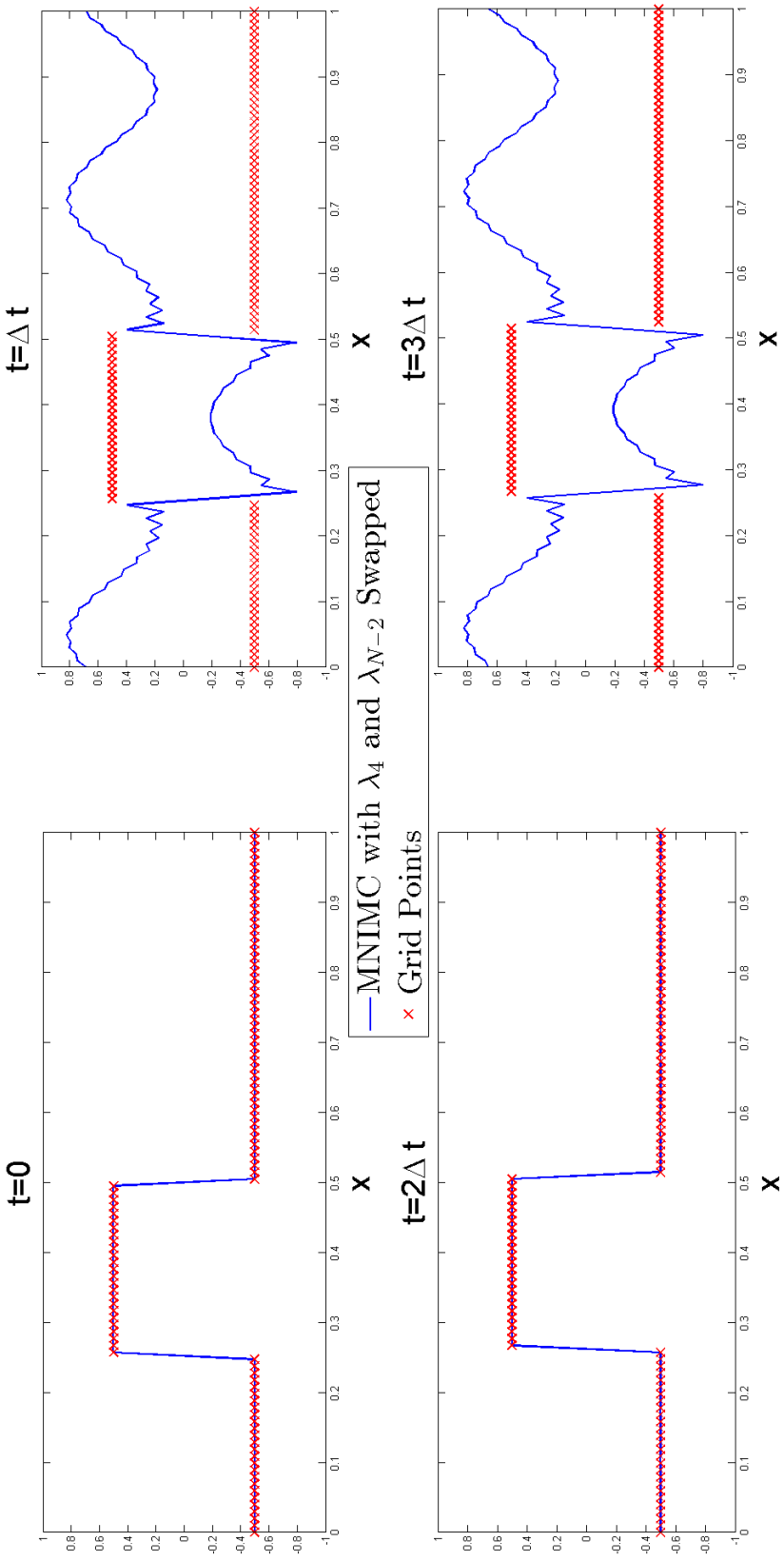


Figure 3.6: *The MNIMC scheme defined in Section 3.7.3, applied to the 1D square function initial condition in (4.28), for* $h = 0.5$, $\mu = 1$, $N_x = 101$ *and* $\Delta t = \frac{1}{202}$, *with the eigenvalues* λ_4 *and* λ_{N_x-2} *swapped to correspond to eigenvectors* $\mathbf{v}_{N_x-2}$ *and* $\mathbf{v}_4$ *respectively. Here we can also see the shifted* $2\Delta t$*-periodic nature of the aliasing error present in the scheme due to the denominator of* h *being equal to two.*

3.10 The effect of numerical dissipation and dispersion on the analysis vector

Numerical dissipation and dispersion are introduced into our considered inverse problem through the forward model M. This Section explores how these errors affect the analysis vector. This is achieved by formulating the analysis vector in terms of the true initial condition, allowing the direct impact of numerically dissipative and/or dispersive eigenvalues of the imperfect scheme, to be seen. Under the conditions of Assumptions 3.2, $\mathcal{M}_{l+1,l} := M$ and $\mathbf{x}_l := \mathbf{U}^l$ for all l, and the cost function in Equation (2.10) becomes,

$$J(\mathbf{x}_0) = \frac{1}{\sigma_o^2} \sum_{l=0}^{L} [\mathbf{y}_l - M^l \mathbf{x}_0]^T [\mathbf{y}_l - M^l \mathbf{x}_0]. \tag{3.67}$$

The aim of the strong constraint 4D-Var data assimilation problem posed in Section 2.3, is to recover the true initial condition $u_0(x)$, sampled at the regularly spaced sample points of the finite difference schemes. This is the vector $\tilde{\mathbf{x}}_0$, defined in Section 3.9. The analysis vector $\mathbf{x}_a$, is the solution to the inverse problem, ie: $\nabla J(\mathbf{x}_a) = 0$. Therefore, we would like to obtain $\mathbf{x}_a = \tilde{\mathbf{x}}_0$. The analysis vector which minimises Equation (3.67) with respect to $\mathbf{x}_0$ is,

$$\mathbf{x}_a = \left[\sum_{k=0}^{L} (M^T M)^k\right]^{-1} \sum_{l=0}^{L} (M^T)^l \mathbf{y}_l = V \left[\sum_{k=0}^{L} (\Lambda^* \Lambda)^k\right]^{-1} \sum_{l=0}^{L} (\Lambda^*)^l V^* \mathbf{y}_l, \tag{3.68}$$

using (3.14). This can be re-written using the 1D DFT,

$$\mathcal{F}(\mathbf{x}_a) = \left[\sum_{k=0}^{L} (\Lambda^* \Lambda)^k\right]^{-1} \left[\sum_{l=0}^{L} (\Lambda^*)^l \mathcal{F}(\mathbf{y}_l)\right] = \left[I_{N_x} + \sum_{k=1}^{L} (\Lambda^* \Lambda)^k\right]^{-1} \left[\sum_{l=0}^{L} (\Lambda^*)^l \mathcal{F}(\mathbf{y}_l)\right]. \tag{3.69}$$

This analysis vector is only affected by observation errors and numerical model errors, as discussed in Section 2.3.

Here the diagonal matrices Λ and Λ^* are known as the *forward* and *adjoint models* [18] respectively, in the 1D DFT basis. In the inverse problem, each set of observations is mapped back in time to $t = 0$, by the adjoint model, M^T. Once the observations have been mapped back to the initial time, they are then summed. This process has the potential to create interference between the corresponding wavenumber components of each set of observations $\mathbf{y}_l$.

Each set of observations $\mathbf{y}_l$ contains observations of the physical system taken at time $l\Delta t$. These observations are taken every Δx in space, creating N_x equally spaced observations. As each set of observations contains the same number of observations, taken at the same spatial locations, this allows the 1D DFT basis to construct the state of the physical system represented by the observations in $\mathbf{y}_l$. The coefficients of the 1D DFT basis in the construction of $\mathbf{y}_l$ are given by $V^* \mathbf{y}_l$. The adjoint model in the

1D DFT basis maps all the coefficients of each set of observations in time back to time $t = 0$ and sums all the coefficients from each set of observations in time, corresponding to the same 1D DFT basis function. It is this summing process which could possibly lead to destructive or constructive interference between sets of observations in time.

Once the sets of observations have been mapped back to time $t = 0$, the result is then normalised with respect to the eigenvalues of the scheme. The set of observations at $t = 0$ acts to regularise the solution of the inverse problem so that the matrix applying the normalisation is always invertible.

Expression (3.69) forms the coefficients of the wavenumber components in the construction of the analysis vector $\mathbf{x}_a$, ie: $\mathbf{x}_a = V\mathcal{F}(\mathbf{x}_a)$. Initially, we wish to consider the analysis vector in the absence of observation errors, so only the effects of numerical model error can be investigated. Observation errors will be re-introduced in Section 4.4. Therefore as discussed in Section 2.3, $\sigma_o^2 = 1$ is chosen in Equation (3.67) together with $\tilde{\mathbf{y}}_l = \tilde{\mathbf{y}}_l$, however this does not affect the formulation of the analysis vector in Equation (3.68).

The following Lemma provides an expression for the analysis vector in terms of the sum of a matrix operation on $\tilde{\mathbf{x}}_0$, $A_L\tilde{\mathbf{x}}_0$ and an aliasing correction term $\boldsymbol{\rho}_L \in \mathbb{R}^{N_x}$, when considering perfect observations $\mathbf{y}_l := \tilde{\mathbf{y}}_l$. The matrix $A_L \in \mathbb{R}^{N_x \times N_x}$ is constructed from the MNIMC scheme and the matrix M implementing the considered numerically dissipative and/or numerically dispersive finite difference scheme.

Lemma 3.13. *Let the assumptions of Lemma 3.12 hold true, allowing $\mathbf{x}_a$ to be stated as in (3.68). Consider perfect observations of the physical system ie: $\mathbf{y}_l = \tilde{\mathbf{y}}_l$ for all $l = 0, \ldots, L$, where $L \in \mathbb{N}_0$ is finite, in the form of (3.54). Then the analysis vector can be expressed as,*

$$\mathbf{x}_a = A_L\tilde{\mathbf{x}}_0 + \boldsymbol{\rho}_L, \tag{3.70}$$

where the model resolution matrix *$A_L \in \mathbb{R}^{N_x \times N_x}$ is such that,*

$$A_L := V\left[\sum_{k=0}^{L}(\Lambda^*\Lambda)^k\right]^{-1}\left[\sum_{l=0}^{L}(\Lambda^*\tilde{\Lambda})^l\right]V^*, \tag{3.71}$$

and $\boldsymbol{\rho}_L \in \mathbb{R}^{N_x}$ is given by,

$$\begin{aligned}\boldsymbol{\rho}_L \;:=\; & V\left[\sum_{k=0}^{L}(\Lambda^*\Lambda)^k\right]^{-1}\left[\left\{\sum_{l=0}^{\frac{L-[L]_b}{b}-1}(\Lambda^*\tilde{\Lambda})^{lb}\right\}\left\{\sum_{y=1}^{b-1}(\Lambda^*)^y V^*\mathbf{r}_y\right\}\right. \\ & \left.+\left(\Lambda^*\tilde{\Lambda}\right)^{L-[L]_b}\left\{\sum_{y=1}^{[L]_b}(\Lambda^*)^y V^*\mathbf{r}_y\right\}\right].\end{aligned} \tag{3.72}$$

Here we consider $\sum_{j=1}^{0} (\Lambda^*)^j V^* \mathbf{r}_j = \mathbf{0}$ *and* $\sum_{l=0}^{-1} \left(\Lambda^* \tilde{\Lambda}\right)^{lb} = 0_{N_x} \in \mathbb{R}^{N_x \times N_x}$ *as we assume* $\mathbf{r}_0 = \mathbf{0}$.

Proof. Consider perfect observations of the 1D linear advection problem, $\mathbf{y}_l = \tilde{\mathbf{y}}_l = \tilde{M}^l \mathbf{x}_0 + \mathbf{r}_l$ for all $l = 0, \ldots, L$, where $L \in \mathbb{N}_0$ is finite. Now consider the analysis vector in Equation (3.68), using these perfect observations. The analysis vector minimises the cost function in Equation (3.67) using $\sigma_o^2 = 1$ and perfect observations. Hence using perfect observations, the analysis vector becomes,

$$\mathbf{x}_a = \left[\sum_{k=0}^{L} (M^T M)^k\right]^{-1} \left[\sum_{l=0}^{L} (M^T \tilde{M})^l \tilde{\mathbf{x}}_0 + \sum_{l=0}^{L} (M^T)^l \mathbf{r}_l\right]. \tag{3.73}$$

Then using the eigenvalue decomposition of M and $\tilde{M}$,

$$\mathbf{x}_a = V \left[\sum_{r=0}^{L} (\Lambda^* \Lambda)^r\right]^{-1} \left\{ \left[\sum_{l=0}^{L} (\Lambda^* \tilde{\Lambda})^l\right] V^* \tilde{\mathbf{x}}_0 + \sum_{l=0}^{L} (\Lambda^*)^l V^* \mathbf{r}_l \right\}.$$

This implies that

$$\mathbf{x}_a = A_L \tilde{\mathbf{x}}_0 + \mathbf{s}_L,$$

where A_L is defined as in Equation (3.71) and,

$$\mathbf{s}_L = V \left[\sum_{r=0}^{L} (\Lambda^* \Lambda)^r\right]^{-1} \left[\sum_{l=0}^{L} (\Lambda^*)^l V^* \mathbf{r}_l\right]. \tag{3.74}$$

The result of Lemma 3.12 in (3.55) gives that $\mathbf{r}_l$ has a shifted $b\Delta t$-periodic nature, where $\mathbf{r}_l = \mathbf{0}$ when $[l]_b = 0$. In order to take advantage of this property when L is finite, (3.74) is rewritten using (3.55) by considering each l in the form $l = sb + [l]_b$, where $s \in \mathbb{N}_0$,

$$\begin{aligned}
\sum_{l=0}^{L} (\Lambda^*)^l V^* \mathbf{r}_l &= \sum_{l=0}^{\frac{L-[L]_b}{b}-1} \sum_{j=0}^{b-1} (\Lambda^*)^{lb+j} V^* \mathbf{r}_{lb+j} + \sum_{j=0}^{[L]_b} (\Lambda^*)^{L-[L]_b+j} V^* \mathbf{r}_{L-[L]_b+j}, \\
&= \left\{ \sum_{l=0}^{\frac{L-[L]_b}{b}-1} \left(\Lambda^* \tilde{\Lambda}\right)^{lb} \right\} \left\{ \sum_{j=1}^{b-1} (\Lambda^*)^j V^* \mathbf{r}_j \right\} \\
&\quad + \left(\Lambda^* \tilde{\Lambda}\right)^{L-[L]_b} \left\{ \sum_{j=1}^{[L]_b} (\Lambda^*)^j V^* \mathbf{r}_j \right\}.
\end{aligned}$$

As a result, $\mathbf{s}_L \equiv \boldsymbol{\rho}_L$ and hence, $\mathbf{x}_a = A_L \tilde{\mathbf{x}}_0 + \boldsymbol{\rho}_L$. □

Expression (3.70) can be viewed as the sum of two analysis vectors, each created by

Equation (3.68), but with two different sets of observations; $\mathbf{y}_l = \tilde{\mathbf{x}}_l$ and $\mathbf{y}_l = \mathbf{r}_l$. As a result, the aliasing error correction terms for the MNIMC scheme $\mathbf{r}_l$, do not play a role in the construction of A_L and are solely found in $\boldsymbol{\rho}_L$. When considering observations of the form $\mathbf{y}_l = \tilde{\mathbf{x}}_l$, these observations are affected by the aliasing errors present in the MNIMC scheme and consequently affect the corresponding analysis vector $A_L\tilde{\mathbf{x}}_0$. The analysis vector $\boldsymbol{\rho}_L$, created by using observations $\mathbf{y}_l = \mathbf{r}_l$, acts as a correction term for the aliasing errors introduced into $A_L\tilde{\mathbf{x}}_0$ by the MNIMC scheme.

The eigenvalues of A_L in (3.71), determine the magnitude and phase change applied to each wavenumber component of $\tilde{\mathbf{x}}_0$, in the construction of $\mathbf{x}_a$. In this way, they can be described as *amplification factors* for the wavenumber components of $\tilde{\mathbf{x}}_0$. Let ν_p be an eigenvalue of A_L such that $\nu_p = |\nu_p|e^{i\kappa_p}$, $\kappa_p \in [-\pi, \pi)$ for $p = 1, \ldots, N_x$. Due to the diagonal structures of Λ and $\tilde{\Lambda}$, ν_p is constructed solely from λ_p and $\tilde{\lambda}_p$, the pth eigenvalues of M and $\tilde{M}$ respectively,

$$\nu_p = \frac{\sum_{l=0}^{L} \overline{\lambda}_p^l \tilde{\lambda}_p^l}{\sum_{k=0}^{L} |\lambda_p|^{2k}}. \tag{3.75}$$

Numerical model error can enter into ν_p via both λ_p and $\tilde{\lambda}_p$. In the case of $\tilde{\lambda}_p$, any error introduced is due to aliasing. As $\overline{\lambda}_p = \lambda_{N_x-p+2}$ and $\overline{\tilde{\lambda}}_p = \tilde{\lambda}_{N_x-p+2}$ for $p = 2, \ldots, N_x$, $\overline{\nu}_p = \nu_{N_x-p+2}$ and $\kappa_p = -\kappa_{N_x-p+2}$ for $p = 2, \ldots, N_x$. Define $\phi_p := \tilde{\theta}_p - \theta_p$, for $p = 1, \ldots, N_x$, as the error in the phase shift applied by λ_p to the corresponding resolvable wavenumber component of the 1D DFT basis. The complex conjugate property of the eigenvalues results in $-\phi_p = \phi_{N_x-p+2}$ for $p = 2, \ldots, N_x$. Then,

$$\nu_p = \begin{cases} 1, & \text{for } |\lambda_p| = 1 \text{ and } \phi_p = 2\pi s, \quad s \in \mathbb{Z}, \\ \frac{1+|\lambda_p|}{1+|\lambda_p|^{L+1}}, & \text{for } |\lambda_p| < 1 \text{ and } \phi_p = 2\pi s, \quad s \in \mathbb{Z}, \\ \frac{1}{L+1}\left|\frac{\sin\left[(L+1)\frac{\phi_p}{2}\right]}{\sin\left[\frac{\phi_p}{2}\right]}\right| e^{i\kappa_p}, & \text{for } |\lambda_p| = 1 \text{ and } \phi_p \neq 2\pi s, \quad s \in \mathbb{Z}, \\ \frac{[1-|\lambda_p|^{L+1}e^{i(L+1)\phi_p}][1-|\lambda_p|^2][1-|\lambda_p|e^{-i\phi_p}]}{[1-|\lambda_p|^{2(L+1)}][1+|\lambda_p|^2-2|\lambda_p|\cos(\phi_p)]}, & \text{for } |\lambda_p| < 1 \text{ and } \phi_p \neq 2\pi s, \quad s \in \mathbb{Z}, \end{cases} \tag{3.76}$$

by the sum of a geometric progression. When $|\lambda_p| = 1$ and $\phi_p \neq 2\pi s$ for some $s \in \mathbb{Z}$,

$$\tan(\kappa_p) = \tan\left(\frac{L\phi_p}{2}\right), \quad \kappa_p \in [-\pi, \pi), \tag{3.77}$$

for $p = 1, \ldots, N_x$.

When λ_p does not introduce numerical model error into the corresponding resolvable wavenumber component, $\nu_p = 1$, so the corresponding resolvable wavenumber component of $\tilde{\mathbf{x}}_0$ is preserved in $\mathbf{x}_a$. A solely numerically dissipative λ_p with respect to the corresponding resolvable wavenumber component, creates an amplification factor that only affects the amplitude of the corresponding resolvable wavenumber component.

In the case of a solely numerically dispersive eigenvalue of M, the amplification factor affects both the phase and amplitude of the corresponding resolvable wavenumber component. The affect on the magnitude is due to interference between the corresponding resolvable wavenumber components making up each set of observations $\mathbf{y}_l$ in the construction of $\mathbf{x}_a$, as discussed in this Section after Equation (3.69). The quantity,

$$\frac{\sin\left[(L+1)\frac{\phi_p}{2}\right]}{\sin\left[\frac{\phi_p}{2}\right]}, \tag{3.78}$$

that makes up a part of ν_p in this instance, is called the Dirichlet Kernel [75].

A numerically dissipative and dispersive eigenvalue of M, creates an amplification factor that is a combination of the numerically dissipative and dispersive amplification factors in Equation (3.76). However, it is not possible to isolate the dissipative and dispersive effects from one another. The magnitude and phase of the spectra of the model resolution matrix for each scheme are plotted in Figures 3.8, 3.10 and 3.12.

The contribution of $\boldsymbol{\rho}_L$ to the analysis vector is not as easy to analyse, but can be reduced by choosing an $\tilde{\mathbf{x}}_0$ that is minimally constructed from unresolvable wavenumber components or by increasing N_x. High wavenumber components are required to resolve small details in a function, such as corners created by discontinuities or rapidly varying functions. Smoother functions are mainly constructed from low wavenumber components [7]. The smoother a function, the higher its regularity, therefore a higher regularity initial condition will reduce $\boldsymbol{\rho}_L$. Here *regularity* is defined as in Definition 3.8. Choosing $h = 1$ leads to $\boldsymbol{\rho}_L = \mathbf{0}$.

In order to understand the effects of numerical dissipation and/or dispersion on the analysis vector, it is not enough to just analyse the error between the analysis vector $\mathbf{x}_a$ and the discrete sample of the true initial condition $\tilde{\mathbf{x}}_0$, using some norm. This does not provide all the information required to assess the quality of the analysis vector. We also need to understand how numerical dissipation and/or dispersion affects the contribution of each wavenumber component, to the numerical solution. This is especially true with initial conditions containing discontinuities.

Therefore, we begin our analysis of the effects of the model resolution matrix on $\tilde{\mathbf{x}}_0$ and the contribution of $\boldsymbol{\rho}_L$ to the analysis vector, by considering a low regularity initial condition in the form of the 1D square function in (4.28). The 1D square function has regularity zero, requiring many high wavenumber components to resolve the edges of the wave. The vector $\tilde{\mathbf{x}}_0$ is then a discrete sample of the 1D square function. The 1D square function allows us to analyse the ability of strong constraint 4D-Var data assimilation, to reconstruct initial conditions that contain unresolvable wavenumber components, in the presence of numerical dissipation and/or dispersion. This tests the effects of numerical dissipation and/or dispersion on strong constraint 4D-Var, in the same way as Durran's "spike test" [6].

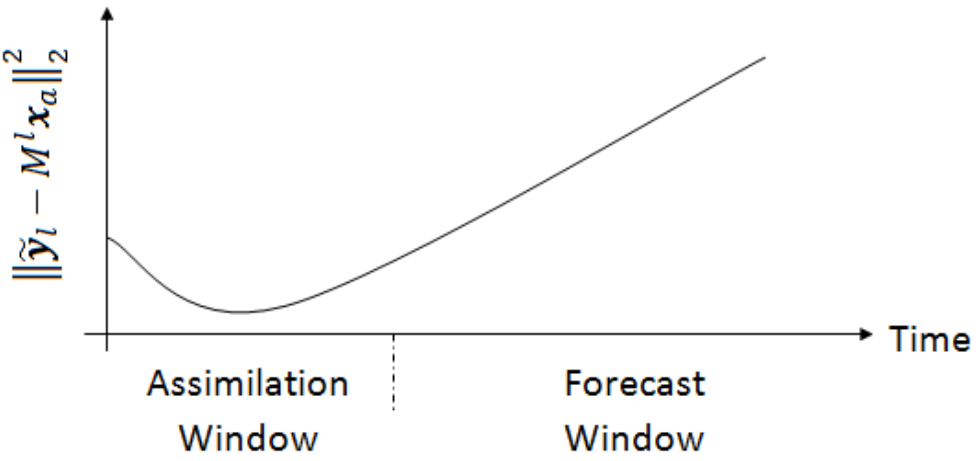
There is also another possible application for this analysis. Strong constraint 4D-

Var data assimilation finds a maximum likelihood estimate [2] for the initial condition of the forward model, used to numerically solve the model equations, for the considered physical system. This estimate is made using observations of the physical system, taken over the assimilation window. Therefore the results of the numerical model, using the analysis vector as an initial condition, provide a good fit to the observations over the assimilation window. The numerical model produces a solution which contains numerical model error. Strong constraint 4D-Var minimises the impact of this error, by constructing the analysis vector so that the error introduced by the numerical model, is implicitly minimised across the assimilation window. The effect is that the point in time when the numerical model is closest to the true state of the physical system, is at some point over the assimilation window. This idea is demonstrated by Figure 3.7(a).

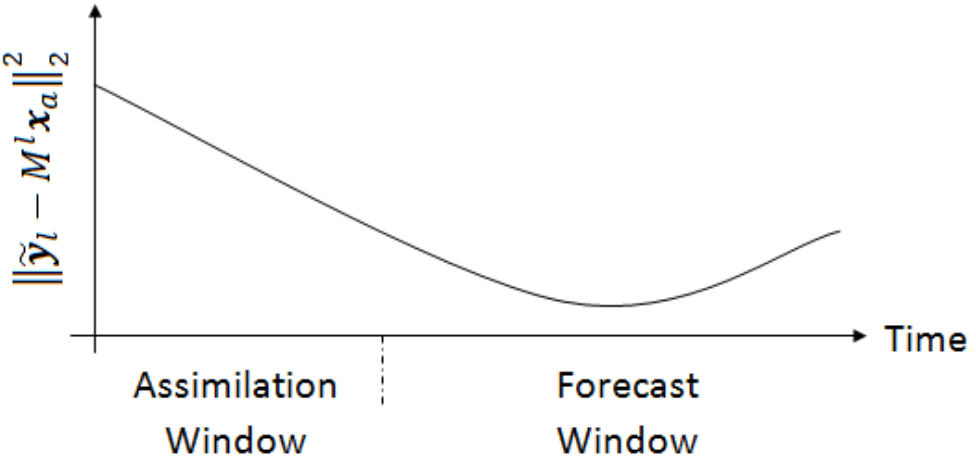
However in applications such as NWP, it is desirable to have the numerical model produce the " 'best' estimate" [2] for the state of the physical system, over the forecast window. Therefore we desire a way to modify strong constraint 4D-Var, such that the effects of numerical model error are minimised across the forecast window, rather than the assimilation window as demonstrated by Figure 3.7(b). This method still needs to make use of the observations taken over the assimilation window. Analysing the structure of the analysis vector may help to design such a method, by allowing the impact of the numerical model on the analysis vector, to be seen directly.

In Sections 3.10.1-3.10.3, the magnitude and phase of the spectra of A_L are analysed for the three schemes, in terms of the real wavenumber components of the solution, together with the result of applying A_L to $\tilde{\mathbf{x}}_0$, for the 1D square function when using $L = 4$. Here we remind the reader that $L + 1$ is the number of sets of observations in time. The corresponding $\boldsymbol{\rho}_L$ and $\mathbf{x}_a$ for the 1D square function are also shown for $L = 4$. The reader is reminded that A_L acts upon all wavenumber components of $\tilde{\mathbf{x}}_0$ through the effects of aliasing. The eigenpair properties of ν_p can be seen through the line of symmetry in the centre of the plots for the magnitude of ν_p and the rotational symmetry in the plots for the phase of ν_p.

We remind the reader here the effects of A_L on the real wavenumber components of $\tilde{\mathbf{x}}_0$, can be seen in the first $\frac{N_x+1}{2}$ (N_x odd) values of p, due to the complex conjugate properties of the eigenvalues of the schemes. This property results in the discontinuity seen in Figure 3.12(b). The magnitude and phase of ν_p are plotted against $(p-1)$ as these are the wavenumbers of the resolvable wavenumber components of the Fourier series for the numerical solution. Increasing p over $p = 1, \ldots, \frac{N_x+1}{2}$, represents increasing the wavenumber of the resolvable real wavenumber component from low to high. The discussions below will make use of this terminology.



(a) Strong constraint 4D-Var data assimilation minimises the effects of numerical model error, over the assimilation window.



(b) Applications such as numerical weather prediction would prefer that strong constraint 4D-Var data assimilation minimise the effects of numerical model error over the forecast window.

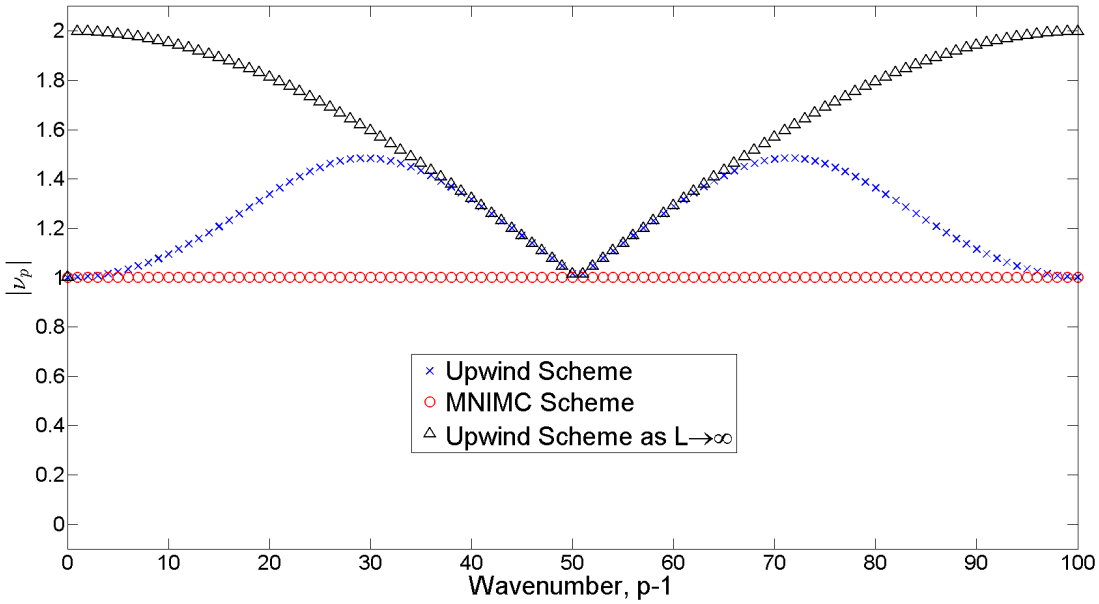
Figure 3.7: *Strong constraint 4D-Var data assimilation minimises the effects of numerical model error, over the assimilation window. Figure 3.7(a) provides a visual representation of this property. It shows that the effects of numerical model error, on the forecast from the analysis vector, increases over the forecast window. Applications such as numerical weather prediction would prefer that the effects of numerical model error on strong constraint 4D-Var data assimilation, be minimised over the forecast window. This idea is represented in Figure 3.7(b).*

3.10.1 The Upwind scheme

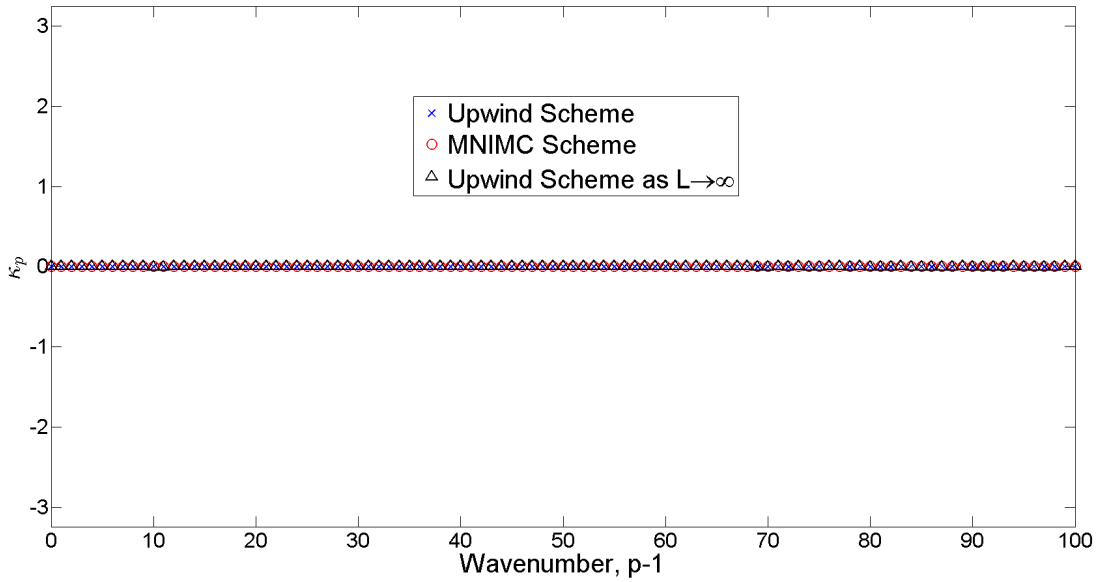
When $h = 0.5$, the Upwind scheme is a numerically dissipative and non-dispersive scheme with respect to the resolvable wavenumber components of the numerical solution, except when $p = 1$. As discussed in Section 3.5.1, this results in the aliasing error introduced by the scheme, being both numerically dissipative and dispersive. These properties of the Upwind scheme and the numerically dispersive aliasing errors introduced by the MNIMC scheme, dictate the oscillations in $A_L\tilde{\mathbf{x}}_0$ and $\boldsymbol{\rho}_L$, compared to $\tilde{\mathbf{x}}_0$ and $\mathbf{0} \in \mathbb{R}^{N_x}$ respectively.

Examining the phase of the eigenvalues of A_L in Figure 3.8(b), we see that all the resolvable wavenumber components of $\tilde{\mathbf{x}}_0$ are propagated with the correct phase speed. As a result, there are no destructive or constructive interference effects affecting the magnitude of the eigenvalues of A_L, in Figure 3.8(a). Examining the magnitude of the eigenvalues of A_L, we see that all but the lowest real resolvable wavenumber components (ie: $p = 1$) of $\tilde{\mathbf{x}}_0$ are amplified by A_L. The greatest amplification effects are experienced by the medium real resolvable wavenumber components. As L increases, the amplification of the lower real resolvable wavenumber components of $\tilde{\mathbf{x}}_0$ increases.

The plots of $A_L\tilde{\mathbf{x}}_0$ and $\boldsymbol{\rho}_L$ in Figure (3.9) demonstrate oscillations at the locations of the discontinuities making up the 1D square function in $\tilde{\mathbf{x}}_0$. The discontinuities are formed from the unresolvable wavenumber components of the 1D square function, so these oscillations represent a failure to propagate the unresolvable wavenumber components of $\tilde{\mathbf{x}}_0$. As $\boldsymbol{\rho}_L$ corrects for the aliasing errors introduced by the MNIMC scheme, this verifies that the oscillations are due to errors in the propagation of the unresolvable wavenumber components of $\tilde{\mathbf{x}}_0$. Adding $\boldsymbol{\rho}_L$ to $A_L\tilde{\mathbf{x}}_0$, removes the effects of aliasing introduced by the MNIMC scheme into $A_L\tilde{\mathbf{x}}_0$, in order to construct $\mathbf{x}_a$ in Figure 3.9(c). This visibly improves the width of the oscillations in $\mathbf{x}_a$ in Figure 3.9(c) when compare to $A_L\tilde{\mathbf{x}}_0$ in Figure 3.9(a). This indicates how important accounting for the effects of aliasing can be. The error in $\mathbf{x}_a$ is solely due to numerical model error introduced by using the Upwind scheme as the forward model. Similar results follow for the remaining schemes in Figures (3.11) and (3.13), with regard to the effects of aliasing errors.

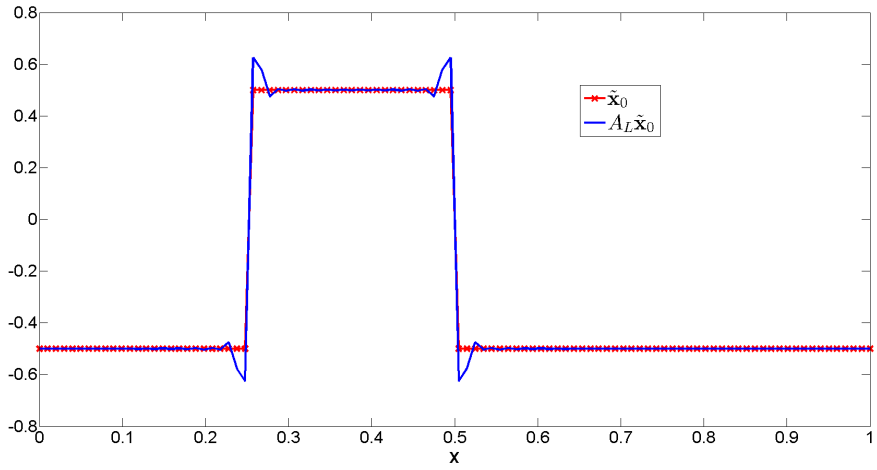


(a) Magnitude of the amplification factors.

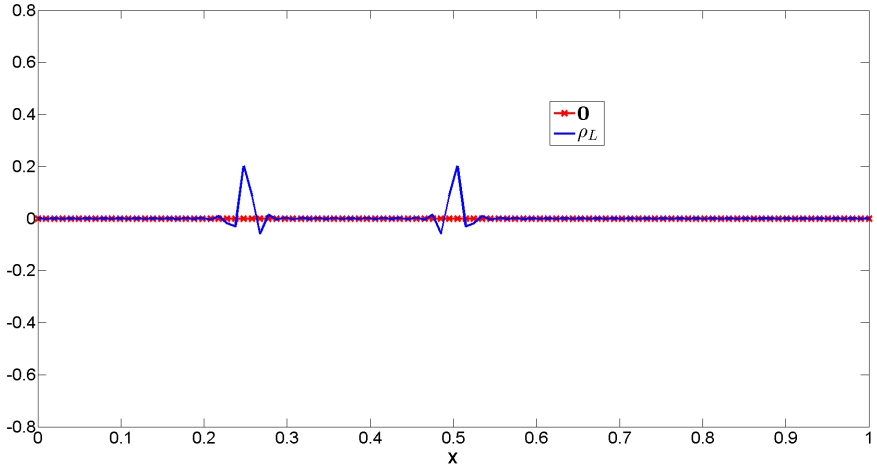


(b) Phase of the amplification factors.

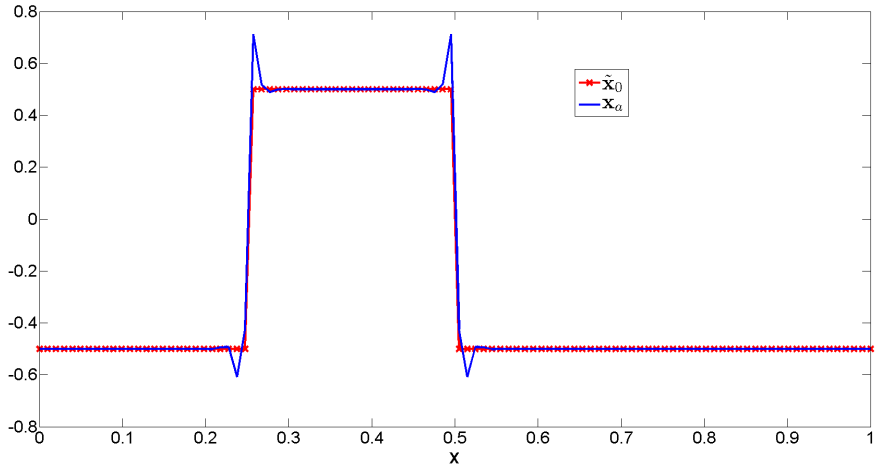
Figure 3.8: *The magnitude and phase of the spectrum of the model resolution matrix, A_L for $L = 4$, together with their limit as $L \to \infty$, for the Upwind scheme when $h = 0.5$, $\mu = 1$ and $N_x = 101$ ($\Delta t = \frac{1}{202}$). The magnitude and phase of the spectrum of A_L for the MNIMC scheme is included for comparison, using the same variables.*



(a) $A_L\tilde{\mathbf{x}}_0$



(b) $\boldsymbol{\rho}_L$



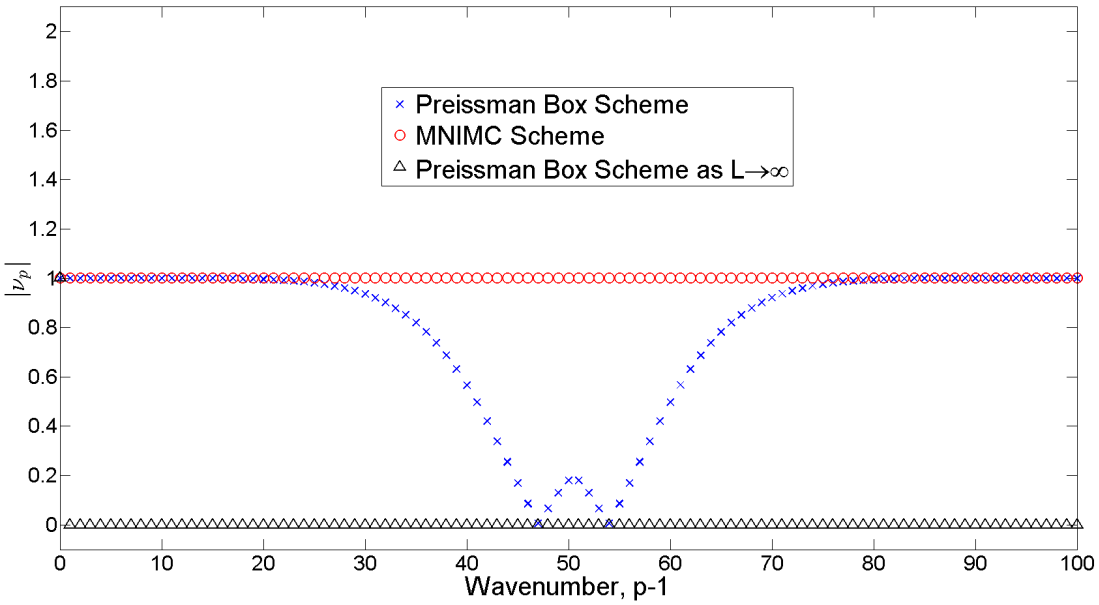
(c) $\mathbf{x}_a = A_L\tilde{\mathbf{x}}_0 + \boldsymbol{\rho}_L$

Figure 3.9: *The analysis vector,* $\mathbf{x}_a = A_L\tilde{\mathbf{x}}_0 + \boldsymbol{\rho}_L$*, for the 1D square function initial condition in (4.28), when using the Upwind scheme and perfect observations,* $\mathbf{y}_l = \tilde{\mathbf{y}}_l = \tilde{\mathbf{x}}_l + \mathbf{r}_l$*, for* $h = 0.5$*,* $\mu = 1$*,* $N_x = 101$ *and* $L = 4$ *(*$\Delta t = \frac{1}{202}$*).*

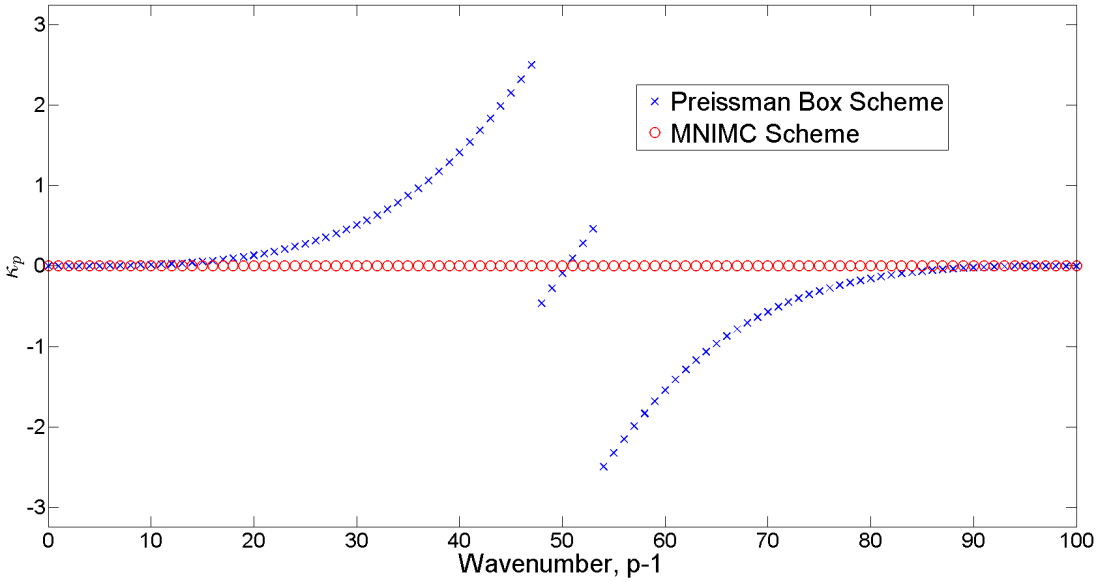
3.10.2 The Preissman Box scheme

The Preissman Box scheme is always numerically non-dissipative with respect to all wavenumber components of the numerical solution, when solving the 1D linear advection problem. When $h = 0.5$ the scheme is numerically dispersive with respect to the resolvable wavenumber components except when $p = 1$, introducing aliasing in the form of numerical dispersion, as discussed in Section 3.5.1. These properties of the Preissman Box scheme and the numerically dispersive aliasing errors introduced by the MNIMC scheme, determine the oscillations in $A_L\tilde{\mathbf{x}}_0$ and $\boldsymbol{\rho}_L$ compared to $\tilde{\mathbf{x}}_0$ and $\mathbf{0} \in \mathbb{R}^{N_x}$, respectively in Figure 3.11. This means that only numerically dispersive effects introduce errors into $A_L\tilde{\mathbf{x}}_0$ and $\boldsymbol{\rho}_L$.

Examining the eigenvalues of A_L in Figure 3.10, we see that the numerically dispersive effects of the schemes affect both the magnitude (Figure 3.10(a)) and phase (Figure 3.10(b)) of the eigenvalues. As there is no numerical dissipation taking place, it is solely the affects of destructive interference between the wavenumber components of sets of observations in time, that is causing the attenuation of the resolvable wavenumber components of $\tilde{\mathbf{x}}_0$. This was discussed in Section 3.10. The amplitude of the lowest resolvable real wavenumber component is the only one not affected ($p = 1$) by destructive interference, as this wavenumber is always correctly propagated by the Preissman Box and MNIMC schemes. In this instance, the low to medium resolvable real wavenumber components experience a small attenuation effect, whilst the medium to high resolvable real wavenumber components experience a much larger attenuation. The highest resolvable real wavenumber components are almost attenuated to zero. As the number of sets of observations in the assimilation window is increased, it is not possible to define a limit for the phase of the eigenvalues of A_L as $L \to \infty$. However Figure 3.10(a) shows that as $L \to \infty$, the magnitude of all eigenvalues of A_L except ν_1, decay to zero. This will be discussed in Section 3.10.5. The effects of destructive interference on the 1D square function initial condition in $\tilde{\mathbf{x}}_0$, can be seen in Figure 3.11. The discussion on the effects of adding $A_L\tilde{\mathbf{x}}_0$ and $\boldsymbol{\rho}_L$ in Figures 3.11(a) and 3.11(b) respectively, to create $\mathbf{x}_a$ in Figure 3.11(c), is similar to that in Section 3.10.1 for the Upwind scheme.

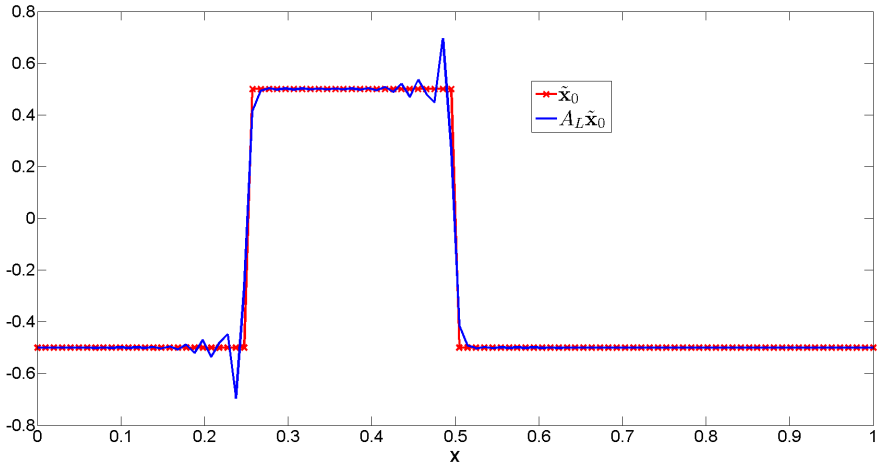


(a) Magnitude of the amplification factors

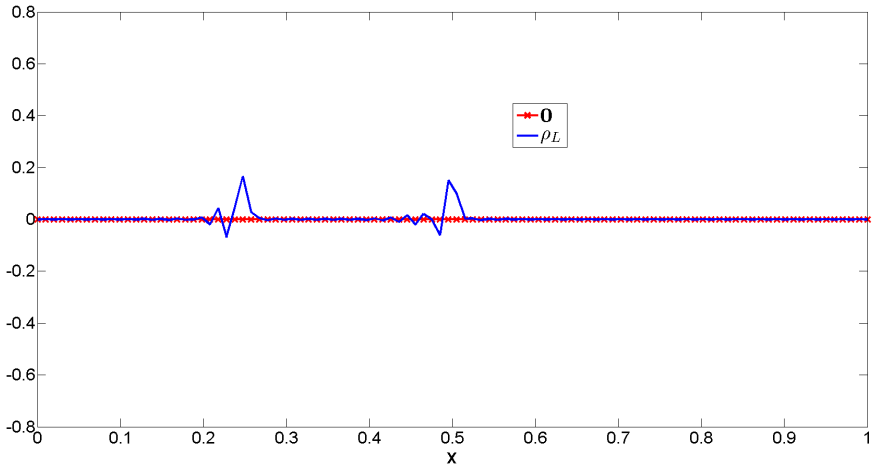


(b) Phase of the amplification factors

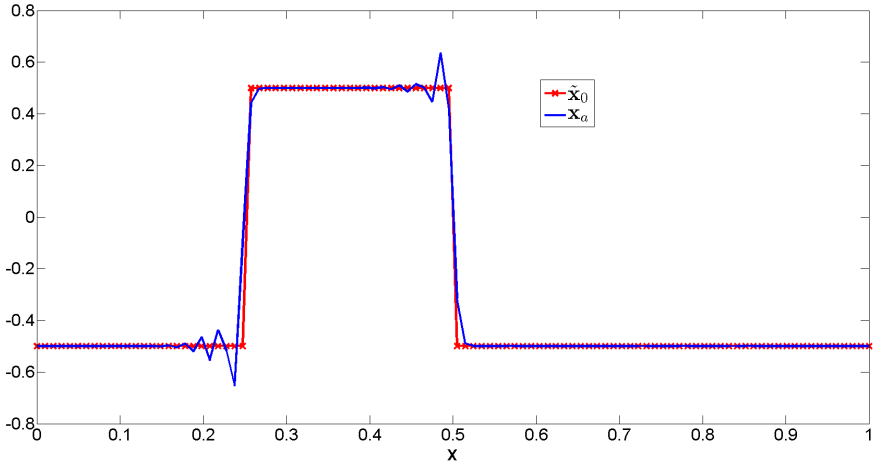
Figure 3.10: *The magnitude and phase of the spectrum of the model resolution matrix, A_L for $L = 4$, together with the limit as $L \to \infty$ for the magnitudes, for the Preissman Box scheme when $h = 0.5$, $\mu = 1$ and $N_x = 101$ ($\Delta t = \frac{1}{202}$). The magnitude and phase of the spectrum of A_L for the MNIMC scheme is included for comparison, using the same variables.*



(a) $A_L\tilde{\mathbf{x}}_0$



(b) $\boldsymbol{\rho}_L$



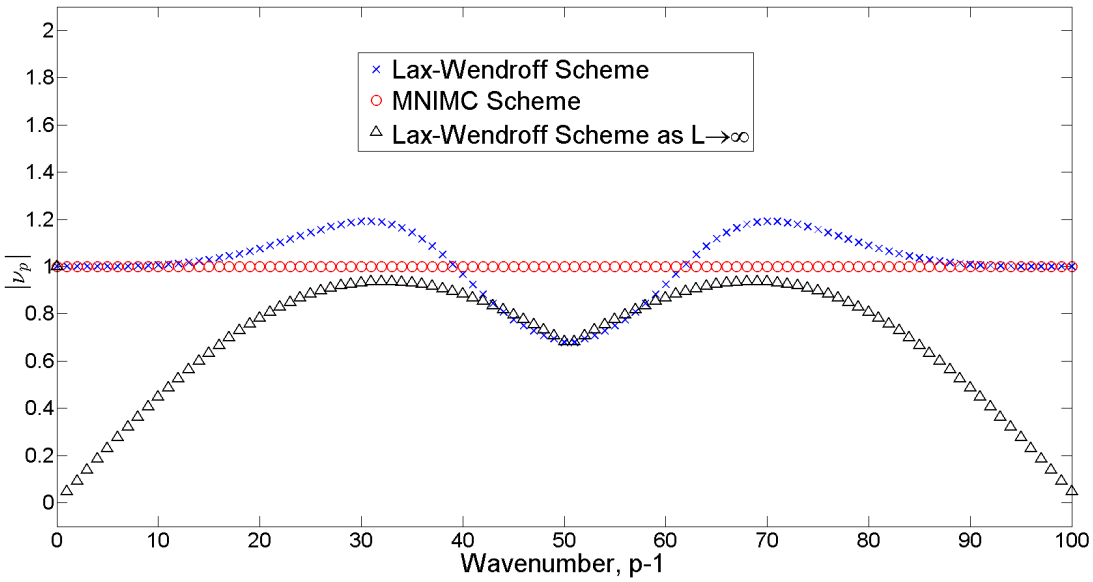
(c) $\mathbf{x}_a = A_L\tilde{\mathbf{x}}_0 + \boldsymbol{\rho}_L$

Figure 3.11: *The analysis vector,* $\mathbf{x}_a = A_L\tilde{\mathbf{x}}_0 + \boldsymbol{\rho}_L$*, for the 1D square function initial condition in (4.28), when using the Preissman Box scheme and perfect observations,* $\mathbf{y}_l = \tilde{\mathbf{y}}_l = \tilde{\mathbf{x}}_l + \mathbf{r}_l$*, for* $h = 0.5$*,* $\mu = 1$*,* $N_x = 101$ *and* $L = 4$ *(*$\Delta t = \frac{1}{202}$*).*

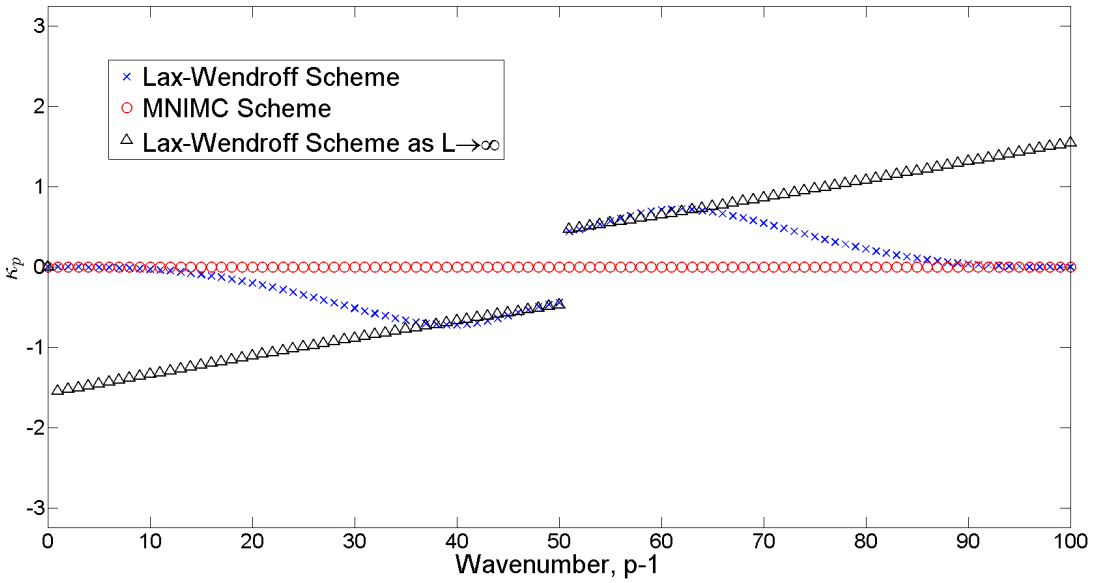
3.10.3 The Lax-Wendroff scheme

When $h = 0.5$, the Lax-Wendroff scheme is both numerically dissipative and dispersive with respect to the resolvable wavenumber components of the numerical solution, except when $p = 1$. As discussed in Section 3.5.1, this results in the aliasing error introduced by the scheme, being both numerically dissipative and dispersive. These properties of the scheme, along with the numerically dispersive aliasing errors introduced by the MNIMC scheme, dictate the oscillations present in $A_L\tilde{\mathbf{x}}_0$ and $\boldsymbol{\rho}_L$ compared to $\tilde{\mathbf{x}}_0$ and $\mathbf{0} \in \mathbb{R}^{N_x}$, respectively.

Examining the eigenvalues of A_L in Figure 3.12, we see that the amplitude of the eigenvalues in 3.12(a) appear to experience a combination of the amplification affects seen in Figure 3.8(a) for the Upwind scheme and the attenuation affects seen in Figure 3.10(a) for the Preissman Box scheme. Comparing the formulation of ν_p for a numerically dissipative and dispersive eigenvalue of a scheme, with that of ν_p for a numerically dissipative and non-dispersive eigenvalue a scheme and ν_p for a numerically non-dissipative and dispersive eigenvalue of a scheme, we see that the former is some combination of the latter two. However, it is not possible to separate the numerically dissipative and dispersive effects in ν_p for a numerically dissipative and dispersive eigenvalue of a scheme. The combination of effects sees the medium and the highest real resolvable wavenumber components of $\tilde{\mathbf{x}}_0$, amplified and attenuated respectively for the Lax-Wendroff scheme, when $L = 4$. The amplification effects seem to balance the attenuation effects so no real resolvable wavenumber components are attenuated to zero. The discussion on the effects of adding $A_L\tilde{\mathbf{x}}_0$ and $\boldsymbol{\rho}_L$ in Figures 3.13(a) and 3.13(b) respectively, to create $\mathbf{x}_a$ in Figure 3.13(c), is similar to that in Section 3.10.1 for the Upwind scheme.

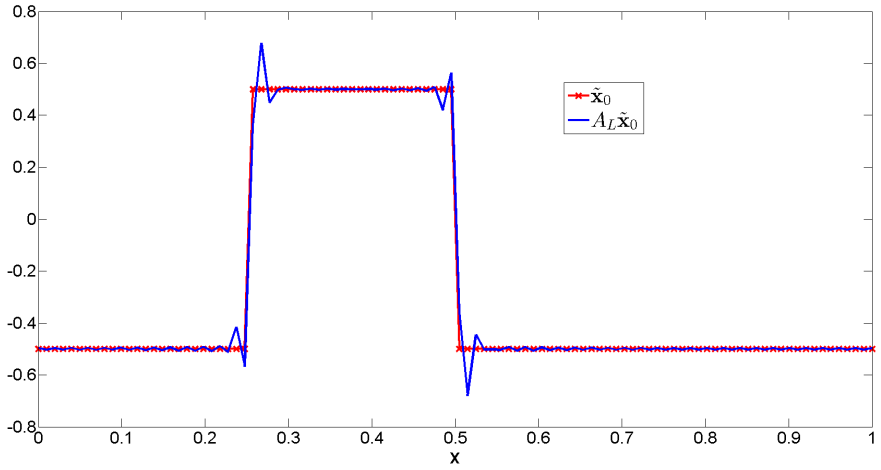


(a) Magnitude of the amplification factors

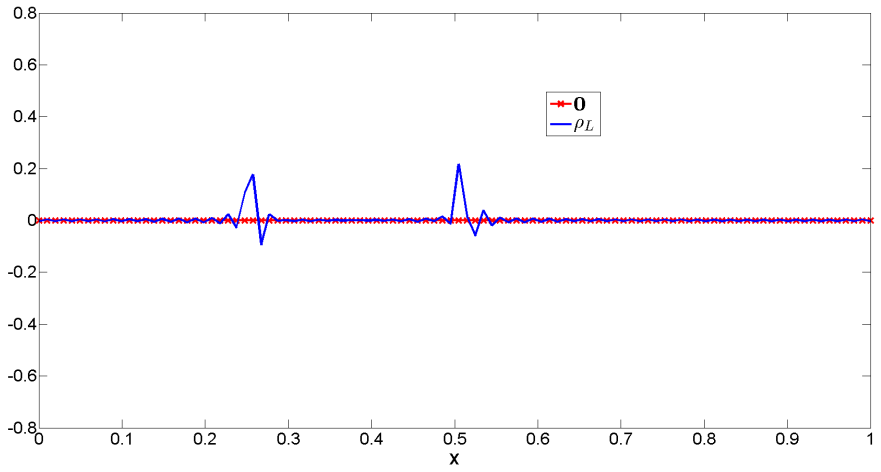


(b) Phase of the amplification factors

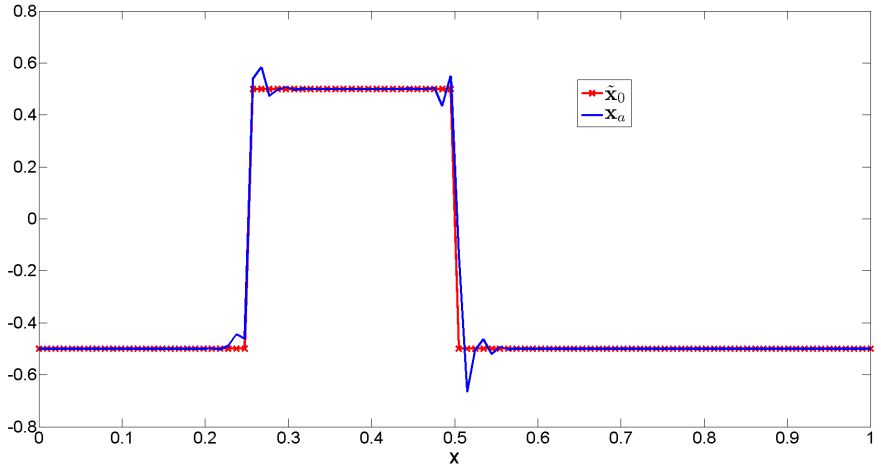
Figure 3.12: *The magnitude and phase of the spectrum of the model resolution matrix, A_L for $L = 4$, together with their limit as $L \to \infty$, for the Lax Wendroff scheme when $h = 0.5$, $\mu = 1$ and $N_x = 101$ ($\Delta t = \frac{1}{202}$). The magnitude and phase of the spectrum of A_L for the MNIMC scheme is included for comparison, using the same variables.*



(a) $A_L\tilde{\mathbf{x}}_0$



(b) $\boldsymbol{\rho}_L$



(c) $\mathbf{x}_a = A_L\tilde{\mathbf{x}}_0 + \boldsymbol{\rho}_L$

Figure 3.13: *The analysis vector,* $\mathbf{x}_a = A_L\tilde{\mathbf{x}}_0 + \boldsymbol{\rho}_L$*, for the 1D square function initial condition in (4.28), when using the Lax-Wendroff scheme and perfect observations,* $\mathbf{y}_l = \tilde{\mathbf{y}}_l = \tilde{\mathbf{x}}_l + \mathbf{r}_l$*, for* $h = 0.5$*,* $\mu = 1$*,* $N_x = 101$ *and* $L = 4$ *(*$\Delta t = \frac{1}{202}$*).*

3.10.4 The MNIMC scheme

The MNIMC scheme is numerically non-dissipative and non-dispersive, with respect to the resolvable wavenumber components of the numerical solution, for any $h \in \mathbb{R}^+$. As a result, $A_L = I_{N_x}$ so $A_L\tilde{\mathbf{x}}_0$ recovers $\tilde{\mathbf{x}}_0$ in Figure 3.14(a). The magnitude and phase of the spectrum of A_L can be seen in Figures 3.8, 3.10 and 3.12. The oscillations in $\boldsymbol{\rho}_L$ in Figure 3.14(b), are due to the aliasing errors in the MNIMC scheme as $h = 0.5$. Despite $A_L\tilde{\mathbf{x}}_0$ recovering $\tilde{\mathbf{x}}_0$, $\boldsymbol{\rho}_L$ is still added to $A_L\tilde{\mathbf{x}}_0$ to create $\mathbf{x}_a$, creating oscillations in $\mathbf{x}_a$ in Figure 3.14(c). When $h \in \mathbb{N}$, the MNIMC scheme does not introduce aliasing errors, so $\boldsymbol{\rho}_L = \mathbf{0}$. Therefore $\mathbf{x}_a = \tilde{\mathbf{x}}_0$ in this instance.

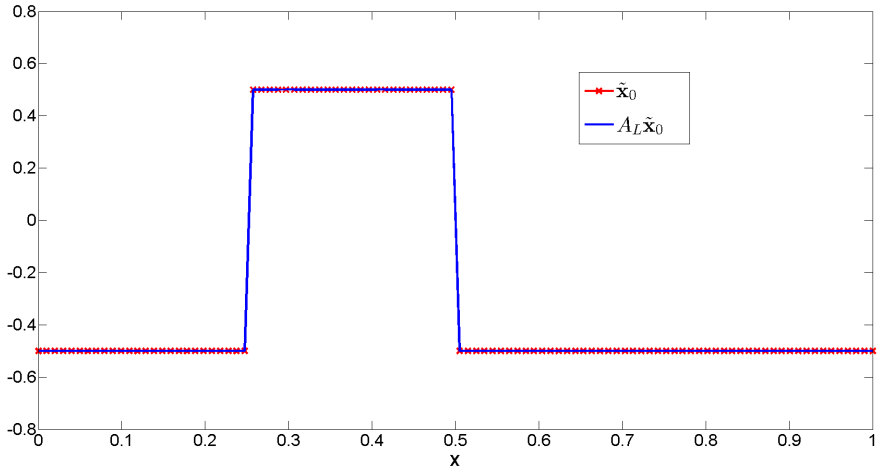
3.10.5 The length of the assimilation window

Another factor that affects the behaviour of numerical model error is the length of the assimilation window. It is important to understand whether the extra time and processing power required to include more observations will yield an improvement in the solution. To understand the behaviour of ν_p for large L, we consider ν_p as $L \to \infty$. As $L \to \infty$,

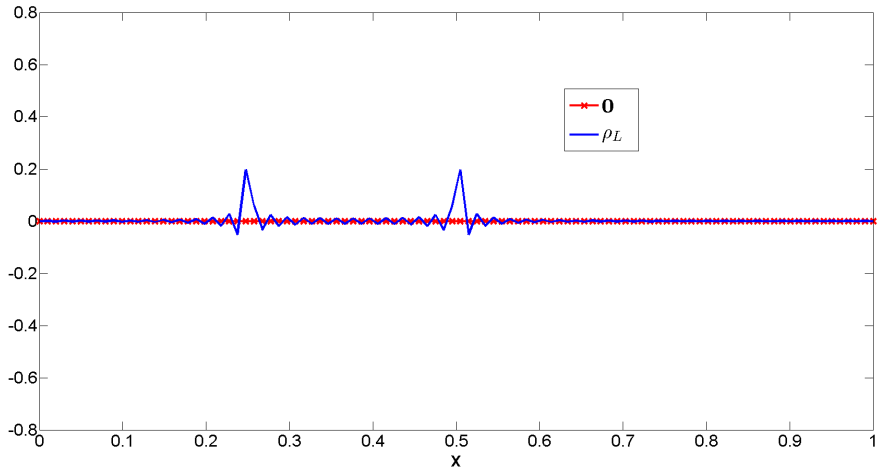
$$\nu_p \to \begin{cases} 1, & \text{for } |\lambda_p| = 1 \text{ and } \phi_p = 2\pi s,\ s \in \mathbb{Z}, \\ 1 + |\lambda_p|, & \text{for } |\lambda_p| < 1 \text{ and } \phi_p = 2\pi s,\ s \in \mathbb{Z}, \\ 0, & \text{for } |\lambda_p| = 1 \text{ and } \phi_p \neq 2\pi s,\ s \in \mathbb{Z}, \\ \frac{(1-|\lambda_p|^2)(1-|\lambda_p|e^{-i\phi_p})}{1+|\lambda_p|^2-2|\lambda_p|\cos(\phi_p)}, & \text{for } |\lambda_p| < 1 \text{ and } \phi_p \neq 2\pi s,\ s \in \mathbb{Z}. \end{cases} \tag{3.79}$$

When $|\lambda_p| \ll 1$, ν_p is very close to its limit for $L \to \infty$, for a relatively small value of L when considering numerically dissipative eigenvalues. This can be seen in Figures 3.8(a) and 3.12(a) where the amplification factors for the highest real resolvable wavenumber components are approaching their limit for $L \to \infty$, when $L = 4$. Hence increasing the length of the assimilation window for the Upwind and Lax-Wendroff schemes, will not affect the contribution of the high resolvable real wavenumber components to the analysis vector and its forecast. The amplification factor for the lower resolvable real wavenumber components can be altered by increasing the length of the assimilation window.

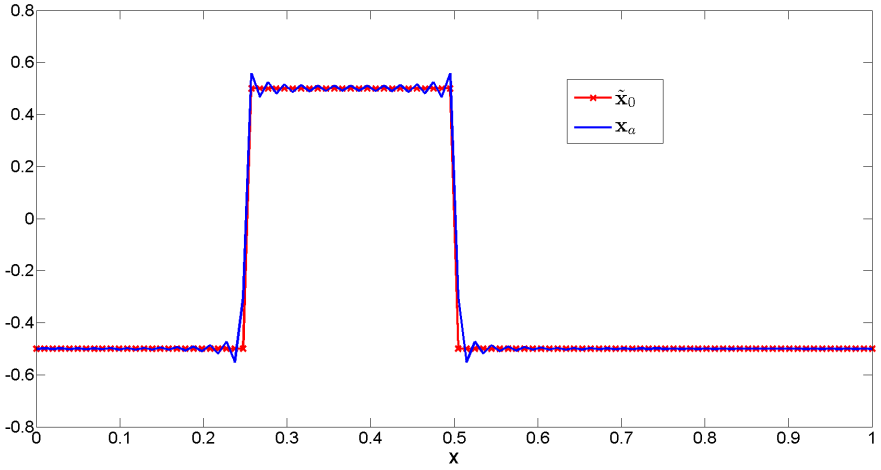
In the case of a numerically non-dissipative and dispersive eigenvalue λ_p, such as those found in the Preissman Box scheme, $\nu_p \to 0$ as $L \to \infty$. This can be seen in Figure 3.10(a). This leads to $A_L\tilde{\mathbf{x}}_0 \to \mathbf{0}$ as $L \to \infty$. Therefore as the length of the length of the assimilation window is increased, by adding extra observations, the contribution of $A_L\tilde{\mathbf{x}}_0$ to $\mathbf{x}_a$ decreases. This shows that as more sets of observations are included in time, destructive interference increases between the corresponding wavenumber components (see Section 3.10), leading to a loss of information in $\mathbf{x}_a$ and its subsequent forecast. Hence for a solely numerically dispersive scheme, increasing the number of sets of observations in the assimilation window does not necessarily improve the accuracy of



(a) $A_L\tilde{\mathbf{x}}_0$



(b) $\boldsymbol{\rho}_L$



(c) $\mathbf{x}_a = A_L\tilde{\mathbf{x}}_0 + \boldsymbol{\rho}_L$

Figure 3.14: *The analysis vector,* $\mathbf{x}_a = A_L\tilde{\mathbf{x}}_0 + \boldsymbol{\rho}_L$*, for the 1D square function initial condition in (4.28), when using the MNIMC scheme and perfect observations,* $\mathbf{y}_l = \tilde{\mathbf{y}}_l = \tilde{\mathbf{x}}_l + \mathbf{r}_l$*, for* $h = 0.5$*,* $\mu = 1$*,* $N_x = 101$ *and* $L = 4$ *(*$\Delta t = \frac{1}{202}$*).*

the analysis vector and its forecast. We are unable to calculate at this point how $\boldsymbol{\rho}_L$ behaves as the number of sets of observations in the assimilation window is increased, as we do not know how $\mathbf{r}_l$ behaves for $l = 1, \ldots, b-1$ as $l \to \infty$. This analysis will be completed in Section 4.3.1.

3.11 Summary

In this Chapter, we have explored the effects of numerical model error, introduced by finite difference schemes, on the analysis vector created through strong constraint 4D-Var data assimilation. The 1D linear advection problem was considered as our physical system. The data assimilation problem set out in Section 2.3 was considered in the absence of all forms of error other than numerical model error. Finite difference schemes were used as forward models to solve the 1D linear advection problem, introducing numerical model error through the approximation of derivatives by finite differences. This error could be classified as numerically dissipative and/or numerically dispersive. Numerical dissipation and numerical dispersion occurred when the magnitude and phase respectively, of a wavenumber component, was incorrectly propagated by the scheme. Metrics were designed to provide a way of measuring the numerically dissipative and dispersive properties of finite difference schemes used to solve the considered 1D linear advection problem.

As all other forms of error had been removed from the problem, we required a way to generate perfect observations of the physical system both numerically and algebraically. Generating perfect observations for the 1D linear advection problem was a challenge and would be for any considered PDE. In the case of the 1D linear advection problem, the form of the solution allows observations to be generated numerically using MATLABs®*circshift* function [74]. However, this is not possible for most other PDEs we could consider. The development of the MNIMC scheme allowed perfect observations to be defined algebraically in terms of the scheme plus an additive correction term to correct for aliasing errors. The scheme was defined to be numerically non-dissipative and non-dispersive with respect to all resolvable wavenumber components of the numerical solution, for all values of the CFL number $h \in \mathbb{R}^+$. However when $h \notin \mathbb{N}$, the scheme introduces aliasing errors in the form of numerical dispersion.

The MNIMC scheme was developed using the analytical solution for the physical system in Fourier series form. It was found that the aliasing errors introduced by the scheme had a shifted periodic nature. This property means that the scheme could be used to generate perfect observations numerically for the physical system as well as algebraically. This asks the question 'is it possible to define such a scheme for other PDEs and will the aliasing errors also have a shifted periodic nature?' This would allow perfect observations to be generated for investigating other PDEs. However, this scheme is not always advantageous as it is computationally expensive.

Using the MNIMC scheme to construct perfect observations algebraically, allowed

the analysis vector to be defined in terms of an amplification matrix acting upon the true initial condition we wish to recover plus an additive term correcting for the accumulative effects of aliasing errors introduced by the MNIMC scheme. This interpretation allowed for the effects of numerical dissipation and/or numerical dispersion in the resolvable wavenumber components of the numerical solution, to be analysed through the eigenvalues of the amplification matrix. By considering the Upwind, Preissman Box, Lax-Wendroff and MNIMC finite difference schemes for $h = 0.5$, this allowed the affects on the analysis vector of numerical dissipation and dispersion on the resolvable wavenumber components of the numerical solution from a scheme, to be viewed individually and in combination.

This analysis revealed that increasing the number of observations in time does not necessarily improve the contribution from real resolvable wavenumber components, whose corresponding eigenvalues in the scheme have a magnitude close to zero and is numerically dissipative with respect to the resolvable wavenumber components of the numerical solution. In the case of the Upwind and Lax-Wendroff schemes, these are the high real resolvable wavenumber components of the analysis vector.

A numerically non-dissipative and dispersive eigenvalue of a scheme with respect to the resolvable wavenumber components, introduces destructive interference between the wavenumber components of sets of observations. This results in a loss of information in the analysis vector. Increasing the number of sets of observations in the assimilation window increases the destructive interference, resulting in an increase in the loss of information. Therefore increasing the number of sets of observations in the assimilation window decreases the accuracy of the analysis vector made using this type of finite difference scheme and hence the forecast made from it. This is counter-intuitive as you would generally expect that providing the 4D-Var process with more observations would increase the accuracy of the analysis vector.

A scheme with a numerically dissipative and dispersive eigenvalue with respect to its corresponding resolvable wavenumber component of the numerical solution, was seen to possess a combination of the effects seen from eigenvalues that were numerically dissipative and non-dispersive and those that were numerically non-dissipative and dispersive with respect to the resolvable wavenumber components of the numerical solution. In this instance, numerical dissipation appeared to reduce the effects of destructive interference, introduced by the numerically dispersive properties of the eigenvalue.

In the next Chapter we continue to investigate the problem considered in this Chapter, but investigate the behaviour of the l_2-norm of the error in the analysis vector, with respect to the number of discretisation points when considering full sets of observations, the number of sets of observations in the assimilation window, the numerically dissipative and dispersive properties of the finite difference scheme and the smoothness of the initial condition.

In the latter half of the next Chapter, we will re-introduce observation errors into the

problem. Observation errors have the greatest impact on the accuracy of the analysis vector [3], so it is important to understand how our considered numerical model error and observation errors behave in combination. To this end, we will also explore the l_2-norm of the error in the analysis vector for this problem, in the same way.

CHAPTER 4

The Effect of Numerical Model Error on the Analysis Vector

One way to measure the impact of numerical model error on 4D-Var data assimilation, is to measure the accuracy of the analysis vector that it creates. The analysis vector is a " 'best' estimate" [2] of the true initial condition, for the system of interest. Applications such as *tomography* and *sonar* and those that use *3D-Var data assimilation*, use the analysis vector directly. Therefore the accuracy of the analysis vector is of great importance for these applications. *Numerical weather prediction (NWP)* makes use of both 3D-Var and 4D-Var data assimilation techniques. 4D-Var data assimilation uses the analysis vector as an initial condition in the numerical model for the system, in order to generate a forecast. Here the accuracy of the forecast over the forecast window is of greatest importance. The accuracy of the forecast is determined by the accuracy of the analysis vector and the numerical model used to generate the forecast. In both applications, the accuracy of the analysis vector plays a role in the accuracy of our desired outcome.

This chapter analyses the behaviour of the l_2-norm of the error in the analysis vector created by strong constraint 4D-Var data assimilation, for the 1D linear advection problem considered in Chapter 3. The advantage of performing this analysis for the 1D linear advection problem is that we know the true initial condition the analysis vector is trying to re-construct. This allows us to quantify the error in the analysis vector and analyse its behaviour [58]. In reality, it generally is not possible to know the true initial condition for the system and hence discover the error present. Therefore it is important to understand the effects of numerical model error on the accuracy of the analysis vector, in problems where the true initial condition can be known. These results can then be translated to more complex problems, to help identify the effects of numerical model error and minimise them.

Analysing the behaviour of the l_2-norm of the error in the analysis vector will allow us to determine whether increasing the number of discretisation points when considering full sets of observations or the number of sets of observations in the assimilation

window, decreases the l_2-norm of the error in the analysis vector. It is important to understand whether the extra computational resources required to make use of this extra data, yields any improvement in the accuracy of the analysis vector. Combining this information with the results on the quality of the analysis vector from Chapter 3, will provide a guide as to the best choice of scheme for our considered strong constraint 4D-Var problem, when considering the 1D linear advection problem.

Initially we consider the error in the analysis vector in the absence of all forms of error, other than numerical model error in the form of numerical dissipation and dispersion, due to the approximation of derivatives by finite differences in the forward model. In the first half of this chapter, we develop two bounds for the error in the analysis vector; one using the truncation error of the considered finite difference scheme and the other using a spectral approach. The aim of such a bound is to determine whether it can be used to characterise the behaviour of the error in the analysis vector due to numerical model error introduced by finite difference schemes. The bound developed through the spectral approach explicitly depends upon the regularity of the true initial condition, the number of discretisation points when considering full sets of observations, the number of sets of observations over the assimilation window and the numerically dissipative and dispersive properties of the finite difference scheme used as the forward model, so we choose to analyse this bound in detail, as we are interested in how these factors affect the error in the analysis vector.

We pose the bound as the sum of six summations and through numerical experiments, determine the dominant summations of the bound which produce its behaviour for each considered scheme. Asymptotic expansions are used to try and characterise the behaviour of these summations analytically. The numerical behaviour of the bound found through the dominant summations, is compared against results from strong constraint 4D-Var data assimilation numerical experiments, for the same schemes and true initial conditions with various regularities. This allows the effectiveness of the bound to characterise the behaviour of the error in the analysis vector, to be assessed.

In the latter half of this Chapter, we will re-introduce observation errors into the problem. Observation errors have the greatest impact on the accuracy of the analysis vector [3], so it is important to understand how our considered numerical model error and observation errors behave in combination. To this end, we will also explore the behaviour of the l_2-norm of the error in the analysis vector for this problem, by developing a bound for this quantity using a similar spectral approach and determining if its suitable for characterising the behaviour of the error.

Once we have completed our analysis, we then move onto a discussion of how the deterministic model error operator for use in the weak constraint 4D-Var problem, can be posed for the 1D linear advection problem.

4.1 Error analysis via the local truncation error

The impact of numerical model error on the accuracy of the analysis vector and its subsequent forecast, can initially be investigated in terms of the local truncation error present in the considered finite difference scheme used as the forward model. Lemma 4.1 describes the error in the analysis vector and its forecast over the forecast window, in terms of the local truncation error of each of our considered finite difference schemes.

Let $\tau_j^l \in \mathbb{R}$ denote the local truncation error, in the lth step of the numerical model implemented by the matrix M, at x_j in space for $l \in \mathbb{N}_0$ and $j = 0, \ldots, N_x - 1$. Then by the consistency of our considered schemes, for sufficiently smooth initial conditions, we have that $\tau_j^l \to 0$ as $\Delta t, \Delta x \to 0$. [14] We now take advantage of this property and use it to show that the l_2-norm of the error in the analysis vector decays to zero as $\Delta t, \Delta x \to 0$.

Lemma 4.1. *Let the conditions in Assumptions 3.2, allowing the matrix M to implement a finite difference scheme for solving the 1D linear advection problem, hold true. Let the CFL number $h \in \mathbb{R}^+$ be constant and valued such that the finite difference scheme implemented by M is convergent. Also, define $\tilde{\mathbf{x}}_0$ as in Section 3.9 and let $u_0(x)$ be sufficiently smooth such that $\tau_j^l \to 0$ as $\Delta x, \Delta t \to 0$ for all $l = 0, \ldots, L$ and $j = 0, \ldots, N_x - 1$.*

Suppose we consider the forecast made from the analysis vector by the forward model M, $s \in \mathbb{N}_0$ time steps after the end of the assimilation window. Then as $\Delta x \to 0$,

(a)

$$\|\tilde{\mathbf{x}}_0 - \mathbf{x}_a\|_2^2 \to 0, \tag{4.1}$$

(b)

$$\left\|\tilde{\mathbf{y}}_{L+s} - M^{L+s}\mathbf{x}_a\right\|_2^2 \to 0. \tag{4.2}$$

Proof. A perfect observation is given by the vector $\tilde{\mathbf{y}}_l$,

$$\tilde{\mathbf{y}}_l = [u(x_0, t^l), \ldots, u(x_{N_x-1}, t^l)]^T, \tag{4.3}$$

for $l \in \mathbb{N}_0$. The truncation error in the lth step of the finite difference scheme at x_j, implemented by the matrix M is defined by,

$$\tau_j^l = \{\tilde{\mathbf{y}}_l\}_{j+1} - \{M\tilde{\mathbf{y}}_{l-1}\}_{j+1},$$

for $j = 0, \ldots, N_x - 1$. Define the vector $\boldsymbol{\tau}_l \in \mathbb{R}^{N_x}$ such that $\boldsymbol{\tau}_l = [\tau_0^l, \ldots, \tau_{N_x-1}^l]^T$, resulting in,

$$\boldsymbol{\tau}_l = \tilde{\mathbf{y}}_l - M\tilde{\mathbf{y}}_{l-1}. \tag{4.4}$$

Then by (4.4),

$$\tilde{\mathbf{y}}_l = \begin{cases} \tilde{\mathbf{y}}_0, & \text{for } l = 0, \\ \sum_{q=1}^{l} M^{l-q}\boldsymbol{\tau}_q + M^l\tilde{\mathbf{y}}_0, & \text{for } l \in \mathbb{N}. \end{cases} \tag{4.5}$$

In order to maintain the numerically dissipative and dispersive properties of the scheme implemented by M, whilst altering the values of Δx and Δt, h must remain constant. When the initial condition sampled in $\tilde{\mathbf{y}}_0$ is sufficiently smooth, $\tau_j^l \to 0$ as $\Delta t, \Delta x \to 0$, keeping h constant.

(a) Consider the error in the analysis vector, using (3.68), together with perfect observations $\mathbf{y}_l := \tilde{\mathbf{y}}_l$. As $\tilde{\mathbf{x}}_0 := \tilde{\mathbf{y}}_0$, substituting in (4.5) results in,

$$\tilde{\mathbf{x}}_0 - \mathbf{x}_a = -\left[\sum_{k=0}^{L}(M^TM)^k\right]^{-1}\sum_{l=1}^{L}(M^T)^l\sum_{q=1}^{l}M^{l-q}\boldsymbol{\tau}_q. \tag{4.6}$$

Taking the l_2-norm we obtain,

$$\begin{aligned} \|\tilde{\mathbf{x}}_0 - \mathbf{x}_a\|_2^2 &\leq \left\{\left\|\left[\sum_{k=0}^{L}(M^TM)^k\right]^{-1}\right\|_2\sum_{l=1}^{L}\sum_{q=1}^{l}\left\|(M^T)^lM^{l-q}\right\|_2\|\boldsymbol{\tau}_q\|_2\right\}^2, \\ &< \left\{\sum_{l=1}^{L}\sum_{q=1}^{l}\|\boldsymbol{\tau}_q\|_2\right\}^2, \end{aligned} \tag{4.7}$$

as $|\lambda_p| \leq 1$ for all $p = 1, \ldots, N_x$. By setting h to be constant, taking $\Delta x \to 0$, results in $\Delta t, \Delta x \to 0$, hence $\tau_j^l \to 0$ for all $j = 0, \ldots, N_x - 1$ and $l = 0, \ldots, L$. Then by the consistency of the numerical model,

$$\|\tilde{\mathbf{x}}_0 - \mathbf{x}_a\|_2^2 \to 0, \quad \text{as } \Delta x \to 0. \tag{4.8}$$

(b) Consider the error in the forecast created by the analysis vector, at time $s\Delta t$ after the end of the assimilation window. Here we again consider perfect observations in order to construct the analysis vector. Then,

$$\tilde{\mathbf{y}}_{L+s} - M^{L+s}\mathbf{x}_a = \sum_{q=1}^{L+s}M^{L+s-q}\boldsymbol{\tau}_q - M^{L+s}\left[\sum_{k=0}^{L}(M^TM)^k\right]^{-1}\sum_{l=1}^{L}(M^T)^l\sum_{q=1}^{l}M^{l-q}\boldsymbol{\tau}_q. \tag{4.9}$$

Taking the l_2-norm we obtain,

$$\begin{aligned}
&\left\|\tilde{\mathbf{y}}_{L+s} - M^{L+s}\mathbf{x}_a\right\|_2^2, \\
\leq\ &\left\{\sum_{q=1}^{L+s} \|M\|_2^{L+s-q} \|\boldsymbol{\tau}_q\|_2 \right. \\
&\left. + \left\|M^{L+s}\right\|_2 \left\|\left[\sum_{k=0}^{L}(M^TM)^k\right]^{-1}\right\|_2 \sum_{l=1}^{L}\sum_{q=1}^{l}\left\|(M^T)^l M^{l-q}\right\|_2 \|\boldsymbol{\tau}_q\|_2 \right\}^2, \\
<\ &\left\{\sum_{q=1}^{L+s} \|\boldsymbol{\tau}_q\|_2 + \sum_{l=1}^{L}\sum_{q=1}^{l} \|\boldsymbol{\tau}_q\|_2 \right\}^2, \qquad (4.10)
\end{aligned}$$

as $|\lambda_p| \leq 1$ for all $p = 1, \ldots, N_x$. By setting h to be constant, taking $\Delta x \to 0$, results in $\Delta t, \Delta x \to 0$, hence $\tau_j^l \to 0$ for all $j = 0, \ldots, N_x - 1$ and $l = 0, \ldots, L + s$. Then by the consistency of the numerical model,

$$\left\|\tilde{\mathbf{y}}_{L+s} - M^{L+s}\mathbf{x}_a\right\|_2^2 \to 0, \ \text{ as } \Delta x \to 0. \qquad (4.11)$$

□

Lemma 4.1 has shown for sufficiently smooth initial conditions and constant h, that the error in the analysis vector and its forecast, decays to zero as $\Delta x \to 0$. As $\Delta x = \frac{1}{N_x}$, this shows that as the number of discretisation points is increased when considering full sets of observations, the error in the analysis vector due to finite difference approximations will decay to zero, for sufficiently smooth initial conditions.

The local truncation error for our considered schemes are; for the Upwind scheme $\tau_j^{l+1} = \mathcal{O}(N_x^{-2})$, for the Preissman Box and Lax-Wendroff schemes $\tau_j^{l+1} = \mathcal{O}(N_x^{-3})$ [7] and for the MNIMC scheme $\tau_j^{l+1} = \mathcal{O}(N_x^{-r})$ where r denotes the regularity of the true initial condition, as defined in Definition 3.8. We derived the consistency property of the MNIMC scheme in Lemma 3.9. It is interesting to note that when $h = 1$, $\tau_j^l = 0$ for all $l \in \mathbb{N}_0$ and $j = 0, \ldots, N_x - 1$, resulting in the bounds in (4.7) and (4.10) being equal to zero. This indicates that there is no error present in the analysis vector or its forecast. When $h = 1$, each of our schemes corresponds to the NIMC scheme. We determined in Section 3.10 that when this scheme is used in strong constraint 4D-Var data assimilation, the analysis vector recovers the discrete sample of the true initial condition $u_0(x)$, so there is no error present in the analysis vector. Therefore our bounds are consistent with this knowledge.

If we consider the order of the truncation errors for each of our considered schemes for $0 < h < 1$, so that the Upwind and Lax-Wendroff schemes are convergent, we find

that,

$$\|\boldsymbol{\tau}_q\|_2 = \begin{cases} \mathcal{O}(N_x^{-\frac{3}{2}}), & \text{for the Upwind scheme,} \\ \mathcal{O}(N_x^{-\frac{5}{2}}), & \text{for the Preissman Box scheme,} \\ \mathcal{O}(N_x^{-\frac{5}{2}}), & \text{for the Lax-Wendroff scheme,} \\ \mathcal{O}(N_x^{-r+\frac{1}{2}}), & \text{for the MNIMC scheme.} \end{cases} \tag{4.12}$$

Substituting these into (4.7) and (4.10), we obtain,

$$\|\tilde{\mathbf{x}}_0 - \mathbf{x}_a\|_2^2 = \begin{cases} \mathcal{O}(L^4 N_x^{-3}), & \text{for the Upwind scheme,} \\ \mathcal{O}(L^4 N_x^{-5}), & \text{for the Preissman Box scheme,} \\ \mathcal{O}(L^4 N_x^{-5}), & \text{for the Lax-Wendroff scheme,} \\ \mathcal{O}(L^4 N_x^{-2r+1}), & \text{for the MNIMC scheme,} \end{cases} \tag{4.13}$$

and

$$\left\|\tilde{\mathbf{y}}_{L+s} - M^{L+s}\mathbf{x}_a\right\|_2^2 = \begin{cases} \mathcal{O}(L^4 N_x^{-3}), & \text{for the Upwind scheme,} \\ \mathcal{O}(L^4 N_x^{-5}), & \text{for the Preissman Box scheme,} \\ \mathcal{O}(L^4 N_x^{-5}), & \text{for the Lax-Wendroff scheme,} \\ \mathcal{O}(L^4 N_x^{-2r+1}), & \text{for the MNIMC scheme,} \end{cases} \tag{4.14}$$

respectively, for sufficiently smooth initial conditions $u_0(x)$. Here we remind the reader that for some functions $f, g : \mathbb{R} \to \mathbb{R}$ such that $x \mapsto f(x), g(x)$, $f(x) = \mathcal{O}(g(x)) \Leftrightarrow$ there exists some constant $C \in \mathbb{R}$ independent of x such that $|f(x)| \leq Cg(x)$ for all $x \in \mathbb{R}$ [14].

Equations (4.13) and (4.14) provide an upper bound for the error in the analysis vector and the error in the forecast $s\Delta t$ through the forecast window, respectively. If they are representative of the behaviour of the respective errors, they show that as the number of discretisation points (N_x) when considering full sets of observations, is increased, the errors decay as was determined in Lemma 4.1. If the bound is tight, then they also show that as the number of sets of observations in the assimilation window (L) is increased, the error in these quantities may increase. This would be counter-intuitive as you would perhaps expect more information over time to increase the accuracy of the analysis vector and consequently the forecast made from it.

We notice that the order of convergence of these bounds with respect to N_x is the same for the Preissman Box and Lax-Wendroff schemes and that they converge to zero faster than the Upwind scheme. This may be due to the numerically dissipative and dispersive properties of the schemes; both the Preissman Box and Lax-Wendroff schemes are numerically dispersive whilst the Upwind scheme is numerically non-dispersive when $h = 0.5$, with respect to the resolvable wavenumber components of the numerical solution. However when $0 < h < 1$ and $h \neq 0.5$, the Upwind scheme is numerically dissipative and dispersive with respect to the resolvable wavenumber components. In this instance, the orders of convergence in (4.12)-(4.14) for the Upwind scheme, remain

unchanged. Therefore we cannot draw any conclusions about how the numerically dissipative and dispersive properties of the schemes, affects the order of convergence of the error, from examining these bounds if they do describe the behaviour of the error. The behaviour of (4.13) and (4.14) with respect to L is identical for each scheme. This is due to L not playing a part in the truncation error of the schemes.

Examining (4.13) and (4.14) for the MNIMC scheme, we see that the smoothness of the true initial condition determines the order of convergence of the bounds with respect to N_x, through its dependency on the regularity (r) of the true initial condition $u_0(x)$. This demonstrates how the smoothness of the initial condition could influence the behaviour of the error in the analysis vector. The truncation error for the Upwind, Preissman Box and Lax-Wendroff schemes does not reveal this information, nor how the numerically dissipative and dispersive properties of the schemes affect it. In the next Section, instead of using the truncation error of the scheme, we use a spectral approach to analyse the error in the analysis vector. This will potentially allow the impact of these properties on the error in the analysis vector, to be observed directly.

4.2 Spectral approach in the absence of observation errors

A spectral approach allows for the error in the analysis vector to be investigated using the eigenvalues of the scheme. Since it is errors in these eigenvalues that are the source of our considered numerical model error, it would seem logical to investigate the error in the analysis vector in this way. We are also able to explicitly observe the effect of the smoothness of the initial condition we wish to recover. The smoothness of the initial condition is given by the regularity of the initial condition, defined in Definition 3.8.

A spectral approach can be used to provide a bound for the l_2-norm of the error in the analysis vector for any regularity initial condition. Provided the bound behaves similarly to the error, it can be analysed to determine the behaviour of the error in the analysis vector. The effectiveness of this analysis is dependent on the tightness of the bound and can be judged through comparing the numerical behaviour of the bound with the same for the l_2-norm of the error in the analysis vector, found through strong constraint 4D-Var data assimilation numerical experiments.

The derivation of such a bound will focus on the eigenvalues of the considered numerically dissipative and/or dispersive finite difference scheme. The result of Lemma 3.13 provides a way to express the analysis vector in terms of the eigenvalues of the considered finite difference scheme and the MNIMC scheme,

$$\|\tilde{\mathbf{x}}_0 - \mathbf{x}_a\|_2^2 = \|(I - A_L)\tilde{\mathbf{x}}_0 - \boldsymbol{\rho}_L\|_2^2 = \|\mathrm{diag}(1-\nu_p)V^*\tilde{\mathbf{x}}_0 - V^*\boldsymbol{\rho}_L\|_2^2 . \tag{4.15}$$

The eigenvalues of the schemes make up $\{\nu_p\}_{p=1}^{N_x}$ and $\boldsymbol{\rho}_L$, as described in (3.76) and (3.72) respectively. Here $\mathrm{diag}(1-\nu_p)$ represents the $N_x \times N_x$ diagonal matrix, where $1-\nu_p$ resides along the main diagonal, in order of increasing p, $p = 1, \ldots, N_x$.

The effect of multiplying $\tilde{\mathbf{x}}_0$ and $\boldsymbol{\rho}_L$ by V^*, is to apply the DFT as discussed in Section 3.3. This identifies the coefficients of the DFT basis in constructing these vectors. The Poisson summation (see Section 3.4.1) allows these coefficients to be considered as a sum of Fourier coefficients, for the initial condition $u_0(x)$, sampled to create the initial condition $\tilde{\mathbf{x}}_0$. As a result in order to create the required bound, a bound on the Fourier coefficients is required.

4.2.1 A bound for the Fourier coefficients

A bound on the Fourier coefficients of a convergent Fourier series is stated in Carslaw [61, p. 269], Henson [66, Theorem 3.5, p. 48] and Briggs and Henson [60, Theorem 6.2, p. 187]. These statements are not accompanied by a proof, however Carslaw [61] and Briggs and Henson [60] provide a sketch proof. This outline has been used to create a proof in Appendix A for the bound on the Fourier coefficients, given in the following Lemma.

Lemma 4.2. *Let $r \in \mathbb{N}_0$ denote the maximum number of times the T-periodic function $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto f(x)$ can be differentiated such that $f^{(\alpha)}(x)$ is continuous and piecewise differentiable over $(0,T)$, for $\alpha = 0, \ldots, r-1$ and $f_{rx}(x)$ is piecewise continuous over $(0,T)$ ie: $f(x)$ has regularity r over $(0,T)$. Also let,*

$$\lim_{x \to 0^+} f^{(\alpha)}(x) = \lim_{x \to T^-} f^{(\alpha)}(x)$$

for $\alpha = 0, \ldots, r-1$ and $f(x)$ be piecewise monotone over $(0,T)$ and $f_{rx}(x)$ be bounded and piecewise monotone over $(0,T)$. Then the coefficients of the Fourier series for $f(x)$, given by $f_k \in \mathbb{C}$, $k \in \mathbb{Z}$, can be bounded such that,

$$|f_k| \leq \begin{cases} D_1, & \text{for } k = 0, \\ \frac{D_2}{|k|^{r+1}}, & \text{for } k \in \mathbb{Z} \setminus \{0\}, \end{cases} \tag{4.16}$$

where $D_1 := v_1 \in \mathbb{R}^+$, the bound on $f(x)$ over $(0,T)$ and $D_2 := \frac{4v_2 s T^r}{(2\pi)^{r+1}}$, where $v_2 \in \mathbb{R}^+$ is the bound on $f^{(r)}(x)$ over $(0,T)$ and s is the number of monotone pieces $f^{(r)}(x)$ can be broken up into on $(0,T)$. This results in D_1 being a constant independent of k, N_x and r and D_2 being a constant independent of k and N_x but dependent on r.

Carslaw's [61] statement of this Lemma requires all derivatives of $f(x)$ up to the $(r-1)$th derivative to be bounded over $(0,T)$ and satisfy Dirichlet's Conditions. We do not require these two conditions in Lemma 4.2. The continuous nature of all derivatives up to the rth derivative of the function ensures that these derivatives are always bounded.

Dirichlet's conditions are not required for the proof as the statement contains the necessary conditions of Theorem 3.1 for $f(x)$ to possess a convergent Fourier series.

The proof in Appendix A allows for the identification of the constants in the bound, where D_1 is independent of k, N_x and r and D_2 is dependent on r, but independent of k and N_x. The proof also adds a small correction to the bound on f_0, compared to the statement of the bound in [61, 66, 60]. This is discussed in detail in Appendix A. The bound is commonly stated over a domain of $(-\pi, \pi)$, see [61, 66, 60]. This results in $D_2 = \frac{2v_2 s}{\pi}$, so the constant is independent of r. It should be noted that when $r = 0$, $v_1 = v_2$. The lemma results in the same bound for f_k and f_{-k} $k \in \mathbb{Z} \setminus \{0\}$, as these coefficients are conjugate pairs.

The bound in (4.16) should be considered as k varies rather than as r varies. It identifies how the Fourier coefficients f_k of a function, decay as the magnitude of the wavenumber k increases [76]. In this sense, it is appropriate to consider the limit of the bound as $|k| \to \infty$. A consequence of the Riemann-Lebesgue lemma [77, Theorem 11.6, p. 313], is that the Fourier coefficients of $f \in L^1([0,T])$, decay as $|k| \to \infty$. Lemma 4.2 provides a bound which can be used to gauge the rate of decay of these Fourier coefficients as $|k| \to \infty$.

Consider the definition of regularity in Definition 3.8. Regularity is defined as the maximum number of times the function $f(x)$ can be differentiated such that it possess a set of properties set out in the definition. Therefore when $k \neq 0$, changing r represents a change of function to one that has a different maximum number of times the function can be differentiated such that the same properties over the same domain hold, with the same values. This means this new function has the same values for v_1 and s over the same domain.

Suppose r is increased in this way and consider the coefficient D_2 in the form $D_2 = \frac{4v_1 T^{r+1}}{T(2\pi)^{r+1}}$ and $k \neq 0$. When $|k|$ is small, there may exist k such that $|k| < \frac{T}{2\pi}$. In this instance $\frac{T}{2\pi|k|} > 1$, so as r increases, the bound on f_k increases. As $|k|$ increases, there may exist $|k|$ such that $|k| = \frac{T}{2\pi}$, in which case $\frac{T}{2\pi|k|} = 1$ and the bound on f_k remains constant with respect to r. Finally, when $|k| > \frac{T}{2\pi}$, $\frac{T}{2\pi|k|} < 1$ so the bound on f_k decreases as r increases. This agrees with our previous discussion in Section 3.10 that smoother functions (higher regularity functions) are constructed mainly from low wavenumber components.

In the case of an initial condition where the function has infinite regularity, such as for a Gaussian function, (4.16) indicates that f_k decays faster than any finite power of k as $|k| \to \infty$ [76]. Boyd [76] states that it is not appropriate to consider (4.16) taking the limit as $r \to \infty$, as the bound was designed to consider f_k as $|k| \to \infty$ for fixed r. If we were to consider the limit as $r \to \infty$ when $\frac{T}{2\pi} < 1$, then for any $k \neq 0$, $\frac{D_2}{|k|^{r+1}} \to 0$ as $|k| \to \infty$. This implies that $f_k = 0$ for all $k \in \mathbb{Z} \setminus \{0\}$, which is generally not true, as is the case for a Gaussian function. In the case of infinite regularity functions, this bound will be considered for large r instead.

4.2.2 A bound for the error in the coefficients found via the 1D DFT

Section 3.4.1 discussed the idea of using the Poisson summation to construct the coefficients of the 1D DFT of a function with a convergent Fourier series. We can use this formulation to determine the error between the coefficient determined by the 1D DFT and coefficient of the Fourier series, for the same resolvable wavenumber component of the function. The bound in (4.16) then allows us to identify a bound for this error. This bound is found stated in Briggs at al. [60, Theorem 6.3, p. 188] and is proved in Henson [66, Theorems 3.6 and 3.7 p. 49-50]. We will find this bound useful for creating a bound on the l_2-norm of the error in the analysis vector, so state it in the following Lemma.

Lemma 4.3. *Let the assumptions of Lemma 4.2 hold true. Additionally let the left- and right-hand derivatives of $f(x)$ with respect to x, exist for all $x \in [0,T]$ and let $x_q = q\Delta x$, where $\Delta x = \frac{T}{N_x}$. Then,*

$$\begin{aligned} \left| \frac{1}{N_x} \sum_{q=1}^{N_x} f(x_{q-1}) e^{\frac{-2\pi i(k-1)(q-1)}{N_x}} - f_{k-1} \right| &\leq \frac{D_3}{N_x^{r+1}}, \quad \text{for } k = 1, \ldots, \left\lfloor \tfrac{N_x}{2} \right\rfloor + 1, \\ \left| \frac{1}{N_x} \sum_{q=1}^{N_x} f(x_{q-1}) e^{\frac{-2\pi i(k-1)(q-1)}{N_x}} - f_{k-1-N_x} \right| &\leq \frac{D_3}{N_x^{r+1}}, \quad \text{for } k = \left\lfloor \tfrac{N_x}{2} \right\rfloor + 2, \ldots, N_x. \end{aligned} \tag{4.17}$$

where,

$$D_3 := \begin{cases} D_2[4 + 2\zeta(2)] + 2v_1 w, & \text{for } r = 0, \\ D_2[2^{r+1} + 2\zeta(r+1)], & \text{for } r \in \mathbb{N}. \end{cases} \tag{4.18}$$

The constants v_1 and s are defined as in Lemma 4.2 and $\zeta(\cdot)$ denotes the Riemann Zeta function. Here $w \in \mathbb{N}$ is the number of sub-domains $[x_j, x_{j+1}]$, $j = 0, \ldots, N_x - 1$, where $u_0(x)$ contains a discontinuity and is assumed to be finite. The constant D_3 is dependent on r, but independent of k and N_x.

The statement of this Lemma has been modified due to the changes in the statement of Lemma 4.2, which it depends upon, and to make use of the notation used for this thesis. An extra case and a small correction have also been added to the proof found in [66]. These are discussed in Appendix A and result in an alternative definition for D_3 when $r = 0$, compared to Henson [66].

Examining Equation (4.17) we see that $\frac{1}{\sqrt{N_x}} \sum_{q=1}^{N_x} f(x_{q-1}) e^{\frac{-2\pi i(k-1)(q-1)}{N_x}}$ is the 1D DFT defined in Section 3.3.1, of the discrete sample of $f(x)$ over $[0,T)$, multiplied by the factor $\frac{1}{\sqrt{N_x}}$ for $k = 1, \ldots, \lfloor \frac{N_x}{2} \rfloor + 1$. So (4.17) provides a bound for the error between the coefficient of a resolvable wavenumber found by the 1D DFT, in comparison to the true Fourier coefficient for that wavenumber component. As previously mentioned, this result is derived using the Poisson summation and the result of Lemma 4.2. This

requires that the Fourier series be convergent, hence the assumption on the left- and right-hand derivatives of $f(x)$ in the statement. The use of Lemma 4.2 means that it is not appropriate to consider this result as $r \to \infty$.

Lemma 4.3 assumes that w does not increase with N_x. In reality, when increasing the number of discretisation points, more subdomains are created. This divides the discontinuities in a sub-domain of the lower resolution discretisation, into several sub-domains in the higher resolution discretisation. As a results, w has the potential to grow with N_x. This means that by increasing the number of discretisation points you may also be increasing the number of grid sub-domains which contain a discontinuity. Hence w may be $\mathcal{O}(N_x)$, altering the outcome of Lemma 4.3 for $r = 0$. This case requires further research.

4.2.3 A bound on the error in the analysis vector

Now Lemmas 4.2 and 4.3 have been established, they can be used to find a bound on the error in the analysis vector, through (4.15).

Lemma 4.4. *Let the assumptions of Lemma 4.3 hold for the function $u_0(x)$ defined in problem (3.1), over its domain $[0, 1)$. Also let the conditions in Assumptions 3.2 hold true, allowing $\mathbf{x}_a$ to be defined as in (3.70) and the MNIMC scheme to be defined as in Definition 3.7. Additionally let $\tilde{\mathbf{x}}_l$ be defined as in Section 3.10, for all $l = 0, \dots, L$. Then,*

$$\begin{aligned} \|\tilde{\mathbf{x}}_0 - \mathbf{x}_a\|_2^2 \;\leq\; & N_x \left\{ |1-\nu_1| D_1 + (|1-\nu_1| + 2\xi_1) \frac{D_3}{N_x^{r+1}} \right\}^2 \\ & + 2N_x \sum_{p=2}^{\frac{N_x+1}{2}} \left\{ |1-\nu_p| \frac{D_2}{(p-1)^{r+1}} + (|1-\nu_p| + 2\xi_p) \frac{D_3}{N_x^{r+1}} \right\}^2, \end{aligned} \tag{4.19}$$

where D_1, D_2 and D_3 are defined as in Lemmas 4.2 and 4.3. Also,

$$\xi_p = \frac{\left| \sum_{l=0}^{\frac{L-[L]_b}{b}-1} \left[|\lambda_p|^b e^{ib\phi_p} \right]^l \right| \left\{ \sum_{y=1}^{b-1} |\lambda_p|^y \right\} + |\lambda_p|^{L-[L]_b} \sum_{y=1}^{[L]_b} |\lambda_p|^y}{\sum_{k=0}^{L} |\lambda_p|^{2k}}. \tag{4.20}$$

Proof. Writing (4.17) for the function $u_0(x)$ in terms of the 1D DFT of $\tilde{\mathbf{x}}_0$ results in,

$$\begin{aligned} \left| \tfrac{1}{\sqrt{N_x}} \mathcal{F}_p(\tilde{\mathbf{x}}_0) - c_{p-1} \right| &\leq \tfrac{D_3}{N_x^{r+1}}, \quad \text{for } p = 1, \dots, \tfrac{N_x+1}{2}, \\ \left| \tfrac{1}{\sqrt{N_x}} \mathcal{F}_p(\tilde{\mathbf{x}}_0) - c_{p-1-N_x} \right| &\leq \tfrac{D_3}{N_x^{r+1}}, \quad \text{for } p = \tfrac{N_x+3}{2}, \dots, N_x. \end{aligned} \tag{4.21}$$

Here N_x is considered odd as this is a requirement for the implementation of the

MNIMC scheme. Equation (3.70) of Lemma 3.13 gives that,

$$\tilde{\mathbf{x}}_0 - \mathbf{x}_a = (I - A_L)\tilde{\mathbf{x}}_0 - \boldsymbol{\rho}_L.$$

Then by taking the l_2-norm and applying the triangle inequality,

$$\|\tilde{\mathbf{x}}_0 - \mathbf{x}_a\|_2^2 = \|(I - A_L)\tilde{\mathbf{x}}_0 - \boldsymbol{\rho}_L\|_2^2 \leq \sum_{p=1}^{N_x} \left\{|1 - \nu_p|\,|\mathcal{F}_p(\tilde{\mathbf{x}}_0)| + |\mathcal{F}_p(\boldsymbol{\rho}_L)|\right\}^2. \tag{4.22}$$

We now require a bound for $|\mathcal{F}_p(\tilde{\mathbf{x}}_0)|$ and $|\mathcal{F}_p(\boldsymbol{\rho}_L)|$ for each p. Consider,

$$\begin{aligned}
&|\mathcal{F}_p(\tilde{\mathbf{x}}_0)| \\
= &\begin{cases} \sqrt{N_x}\left|\frac{1}{\sqrt{N_x}}\mathcal{F}_p(\tilde{\mathbf{x}}_0) - c_{p-1} + c_{p-1}\right|, & \text{for } p = 1, \ldots, \frac{N_x+1}{2}, \\ \sqrt{N_x}\left|\frac{1}{\sqrt{N_x}}\mathcal{F}_p(\tilde{\mathbf{x}}_0) - c_{p-1-N_x} + c_{p-1-N_x}\right|, & \text{for } p = \frac{N_x+3}{2}, \ldots, N_x, \end{cases} \qquad (4.23) \\
\leq &\begin{cases} \sqrt{N_x}\left\{|c_{p-1}| + \left|\frac{1}{\sqrt{N_x}}\mathcal{F}_p(\tilde{\mathbf{x}}_0) - c_{p-1}\right|\right\}, & \text{for } p = 1, \ldots, \frac{N_x+1}{2}, \\ \sqrt{N_x}\left\{|c_{p-1-N_x}| + \left|\frac{1}{\sqrt{N_x}}\mathcal{F}_p(\tilde{\mathbf{x}}_0) - c_{p-1-N_x}\right|\right\}, & \text{for } p = \frac{N_x+3}{2}, \ldots, N_x, \end{cases} \\
\leq &\begin{cases} \sqrt{N_x}\left\{D_1 + \frac{D_3}{N_x^{r+1}}\right\}, & \text{for } p = 1, \\ \sqrt{N_x}\left\{\frac{D_2}{|p-1|^{r+1}} + \frac{D_3}{N_x^{r+1}}\right\}, & \text{for } p = 2, \ldots, \frac{N_x+1}{2}, \\ \sqrt{N_x}\left\{\frac{D_2}{|p-1-N_x|^{r+1}} + \frac{D_3}{N_x^{r+1}}\right\}, & \text{for } p = \frac{N_x+3}{2}, \ldots, N_x, \end{cases} \qquad (4.24) \\
&\text{by (4.16) and (4.17),}
\end{aligned}$$

Now consider $|\mathcal{F}_p(\boldsymbol{\rho}_L)|$ where $\boldsymbol{\rho}_L$ is defined in (3.72). By the triangle inequality, a bound is given in terms of $|\mathcal{F}_p(\mathbf{r}_l)|$ for finite L,

$$|\mathcal{F}_p(\boldsymbol{\rho}_L)| \leq \frac{\left|\sum_{l=0}^{\frac{L-[L]_b}{b}-1}\left(\overline{\lambda}_p\tilde{\lambda}_p\right)^{lb}\right|\left\{\sum_{j=1}^{b-1}|\lambda_p|^j|\mathcal{F}_p(\mathbf{r}_j)|\right\} + |\lambda_p|^{L-[L]_b}\left\{\sum_{j=1}^{[L]_b}|\lambda_p|^j|\mathcal{F}_p(\mathbf{r}_j)|\right\}}{\sum_{r=0}^{L}|\lambda_p|^{2r}}, \tag{4.25}$$

for $p = 1, \ldots, N_x$.

Consider $|\mathcal{F}_p(\mathbf{r}_j)|$ for $1 \leq j \leq b-1$, using (3.54), add and subtract the relevant

continuous Fourier coefficient and apply (4.17),

$$
\begin{aligned}
& |\mathcal{F}_p(\mathbf{r}_j)| \\
= \; & |\mathcal{F}_p(\tilde{\mathbf{y}}_j) - \tilde{\lambda}_p^j \mathcal{F}_p(\tilde{\mathbf{x}}_0)|, \\
= \; & \begin{cases} \sqrt{N_x} \left| \frac{1}{\sqrt{N_x}} \mathcal{F}_p(\tilde{\mathbf{y}}_j) - c_{p-1} e^{\frac{-2\pi i(p-1)jh}{N_x}} + c_{p-1} e^{\frac{-2\pi i(p-1)jh}{N_x}} - \tilde{\lambda}_p^j \frac{1}{\sqrt{N_x}} \mathcal{F}_p(\tilde{\mathbf{x}}_0) \right|, \\ \text{for } p = 1, \ldots, \frac{N_x+1}{2}, \\ \sqrt{N_x} \left| \frac{1}{\sqrt{N_x}} \mathcal{F}_p(\tilde{\mathbf{y}}_j) - c_{p-1-N_x} e^{\frac{-2\pi i(p-1-N_x)jh}{N_x}} + c_{p-1-N_x} e^{\frac{-2\pi i(p-1-N_x)jh}{N_x}} \right. \\ \left. - \tilde{\lambda}_p^j \frac{1}{\sqrt{N_x}} \mathcal{F}_p(\tilde{\mathbf{x}}_0) \right|, \\ \text{for } p = \frac{N_x+3}{2}, \ldots, N_x, \end{cases} \\
\leq \; & \begin{cases} \sqrt{N_x} \left\{ \left| \frac{1}{\sqrt{N_x}} \mathcal{F}_p(\tilde{\mathbf{y}}_j) - c_{p-1} e^{\frac{-2\pi i(p-1)jh}{N_x}} \right| + \left| \frac{1}{\sqrt{N_x}} \mathcal{F}_p(\tilde{\mathbf{x}}_0) - c_{p-1} \right| \right\}, \\ \text{for } p = 1, \ldots, \frac{N_x+1}{2}, \\ \sqrt{N_x} \left\{ \left| \frac{1}{\sqrt{N_x}} \mathcal{F}_p(\tilde{\mathbf{y}}_j) - c_{p-1-N_x} e^{\frac{-2\pi i(p-1-N_x)jh}{N_x}} \right| + \left| \frac{1}{\sqrt{N_x}} \mathcal{F}_p(\tilde{\mathbf{x}}_0) - c_{p-1-N_x} \right| \right\}, \\ \text{for } p = \frac{N_x+3}{2}, \ldots, N_x, \end{cases} \\
\leq \; & \frac{2\sqrt{N_x} D_3}{N_x^{r+1}}, \qquad (4.26)
\end{aligned}
$$

The bound on the error in $\mathcal{F}_p(\tilde{\mathbf{y}}_l)$ also uses the same D_3 coefficient, because $\mathcal{F}_p(\tilde{\mathbf{y}}_l)$ is the 1D DFT of $u(x, l\Delta t) = u(x - \mu l \Delta t, 0) = u_0([x - \mu l \Delta t]_1)$, so has the same bounds and number of monotone pieces as $u_0(x)$ over $[0, 1)$.

Let ξ_p be defined as in the statement of the Lemma for $p = 1, \ldots, N_x$. Notice that as $\overline{\lambda}_p = \lambda_{N_x-p+2}$ for $p = 2, \ldots, N_x$, similarly $\xi_p = \xi_{N_x-p+2}$ for $p = 2, \ldots, N_x$. Combining (4.25) with (4.26) results in,

$$
|\mathcal{F}_p(\boldsymbol{\rho}_L)| \leq \frac{2\sqrt{N_x} D_3 \xi_p}{N_x^{r+1}}. \qquad (4.27)
$$

Then combining (4.22), (4.24) and (4.27) and as $|1 - \nu_p| = |1 - \nu_{N_x-p+2}|$ and $\xi_p = \xi_{N_x-p+2}$ for $p = 2, \ldots, N_x$, for $r \in \mathbb{N}_0$, results in (4.19). □

Lemma 4.4 provides a bound for the l_2-norm of the error in the analysis vector. It is explicitly dependent on the regularity of the initial condition $u_0(x)$ over $(0, 1)$ denoted by r, the dissipative and dispersive properties of the imperfect finite difference scheme via ν_p, the number of discretisation points in space given by N_x when considering full sets of observations and the number of sets of observations taken over the assimilation window L. As Lemma 4.4 makes use of the results from Lemmas 4.2 and 4.3, it is not appropriate to consider (4.19) at the limit $r \to \infty$. Instead large r will be considered in this instance.

In Section 3.10, we determined that by choosing $h = 1$ for any of our schemes, the

error in the analysis vector is zero. This is equivalent to choosing the NIMC scheme. Examining the bound in (4.19), we find that when $h = 1$, this bound is also zero. This means that our bound is consistent with this property and also concurs with our analysis of the error in the analysis vector for $h = 1$, via the local truncation error in Section 4.1.

In the case of the Upwind, Preissman Box and Lax-Wendroff schemes, $\nu_1 = 1$. Hence the terms relating to ν_1 in the bound in (4.19), are zero. In the case of the MNIMC scheme, $\nu_p = 1$ for all $p = 1, \ldots, N_x$ as $A_L = I$, so the only contribution to the error in this case is from aliasing errors. The bound is consistent with this as only the bound on the aliasing error remains.

Now we have a bound on the error in the analysis vector, we want to determine how it behaves with respect to each variable it is dependent on. Through understanding its behaviour and by comparing it to the behaviour of the actual error in the analysis vector determined through numerical experiments, we can determine its suitability for characterising the error in the analysis vector. If the bound possesses similar properties to that of the error, the process of deriving the bound could be used in problems where the error cannot be determined exactly, to provide a way to determine the behaviour of the error.

4.3 Analysis of the bound

Initially, to demonstrate the effectiveness of the bound derived in Lemma 4.4, it is plotted against $\|\tilde{\mathbf{x}}_0 - \mathbf{x}_a\|_2^2$ obtained through strong constraint 4D-Var numerical experiments. As discussed in Section 3.6.4, we choose $h = 0.5$ and $\mu = 1$. These experiments were performed using the Upwind, Preissman Box, Lax-Wendroff and MNIMC finite difference schemes as forward models for solving the 1D linear advection problem in (3.1) and three different functions for $u_0(x)$ over $[0, 1)$. These initial conditions all have differing regularities. These functions are;

- the 1D square function ($r = 0$),

$$u_0(x) = \begin{cases} -\frac{1}{2}, & \text{for } x \in \left[0, \frac{1}{4}\right) \bigcup \left(\frac{1}{2}, 1\right), \\ \frac{1}{2}, & \text{for } x \in \left[\frac{1}{4}, \frac{1}{2}\right]. \end{cases} \tag{4.28}$$

  Here, $v_1 = v_2 = \frac{1}{2}$ and $s = 3$, so $D_1 = \frac{1}{2}$, $D_2 = \frac{3}{\pi}$ and $D_3 = D_2[4 + 2\zeta(2)] + w$, where,

$$w = \begin{cases} 2, & \text{for } \frac{N_x}{4}, \frac{N_x}{2} \notin \mathbb{N}, \\ 3, & \text{for } \frac{N_x}{4} \in \mathbb{N} \text{ and } \frac{N_x}{2} \notin \mathbb{N}, \\ 3, & \text{for } \frac{N_x}{4} \notin \mathbb{N} \text{ and } \frac{N_x}{2} \in \mathbb{N}, \\ 4, & \text{for } \frac{N_x}{4}, \frac{N_x}{2} \in \mathbb{N}, \end{cases} \tag{4.29}$$

  (As the MNIMC scheme requires that N_x be odd, the following experiments will be performed for odd N_x, resulting in $w = 2$.)

- the triangular function ($r = 1$),

$$u_0(x) = \begin{cases} -\frac{1}{2}, & \text{for } x \in \left[0, \frac{1}{4}\right) \bigcup \left(\frac{1}{2}, 1\right), \\ 8x - \frac{5}{2}, & \text{for } x \in \left[\frac{1}{4}, \frac{3}{8}\right], \\ -8x + \frac{7}{2}, & \text{for } x \in \left(\frac{3}{8}, \frac{1}{2}\right]. \end{cases} \tag{4.30}$$

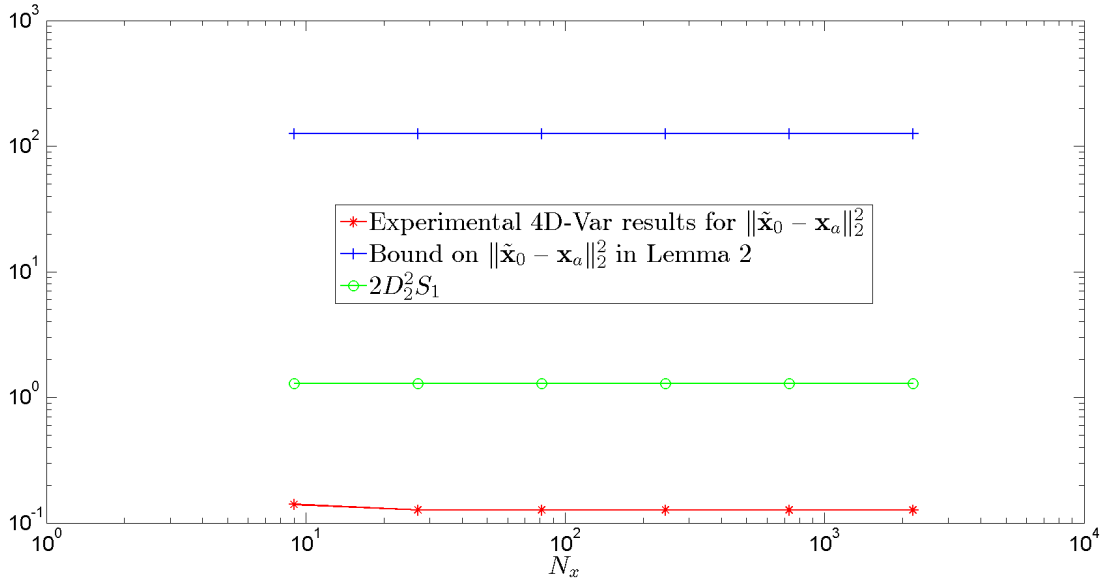
  Here, $v_1 = \frac{1}{2}$, $v_2 = 8$ and $s = 4$, so $D_1 = \frac{1}{2}$, $D_2 = \frac{32}{\pi^2}$ and $D_3 = \frac{64(2+\zeta(2))}{\pi^2} = \frac{128}{\pi^2} + \frac{32}{3}$.

- the 1D Gaussian function $\mathcal{N}\left(\frac{1}{2}, \frac{1}{100}\right)$ ($r \gg 1$),

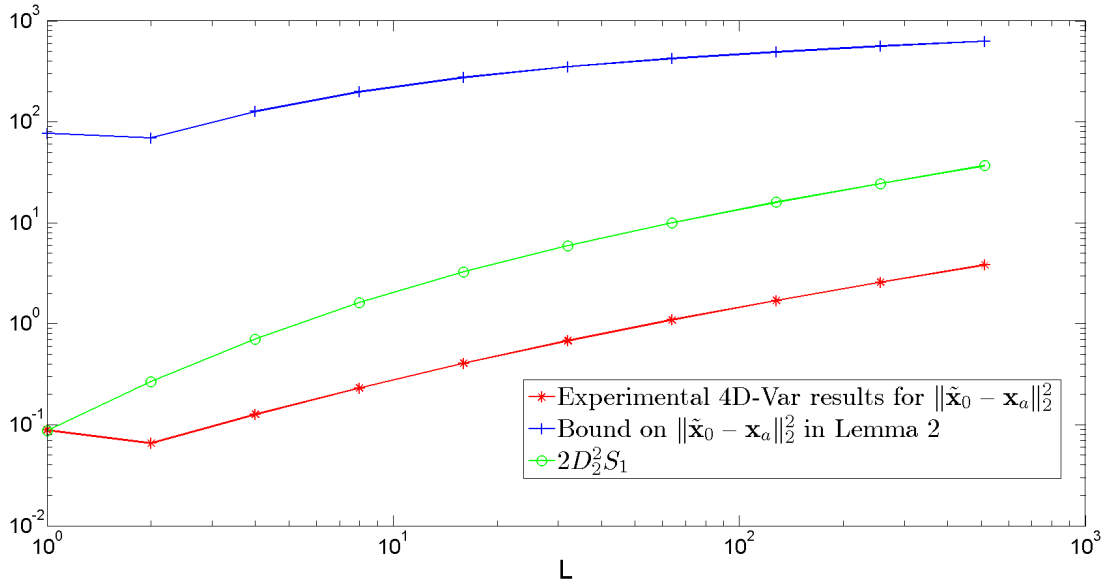
$$u_0(x) = \frac{10}{\sqrt{2\pi}} e^{-50(x-\frac{1}{2})^2}. \tag{4.31}$$

  Here $v_1 = \frac{10}{\sqrt{2\pi}}$. Since we are considering the 1D Gaussian function for large r, we need a way of choosing r such that the bound in (4.19) is sufficiently large, but not infinite. Sections 4.3.3 and 4.3.6 show that if we choose $r = 3$ for the Upwind scheme and $r = 5$ for the Preissman Box and Lax-Wendroff schemes, the numerical order of convergence to zero for the l_2-norm of the error in the analysis vector, will have reached its saturated rate of decay. These values of r are sufficient to obtain the orders of convergence for the error, for large r. A value for v_2 can then be calculated by bounding $u_0^{(3)}(x)$ for the Upwind and $u_0^{(5)}(x)$ for the Preissman Box and Law-Wendroff schemes, using Hermite functions [78].

The strong constraint 4D-Var numerical experiments were executed using the built-in *PCG* method in MATLAB®[74], a zero first guess and a tolerance of 10^{-10} on the relative residual which was reached during numerical experiments, to minimise the cost function. The number of observations is given by $N_x(L+1)$ so as N_x and L are increased the number of observations increases. It should be noted here that in the following experiments, the order of convergence with respect to either N_x or L is found for constant h. This results in the length of the assimilation window varying in time, $L\Delta t = \frac{Lh}{\mu N_x}$, as N_x and L are varied. Also as $\Delta t = \frac{h}{\mu N_x}$, keeping h constant whilst changing N_x results in Δt varying for each value of N_x. This together with the fact that each set of observations $\tilde{\mathbf{y}}_l$ contains observations taken at every grid point in space, means that increasing N_x increases the density of observations in space and time.

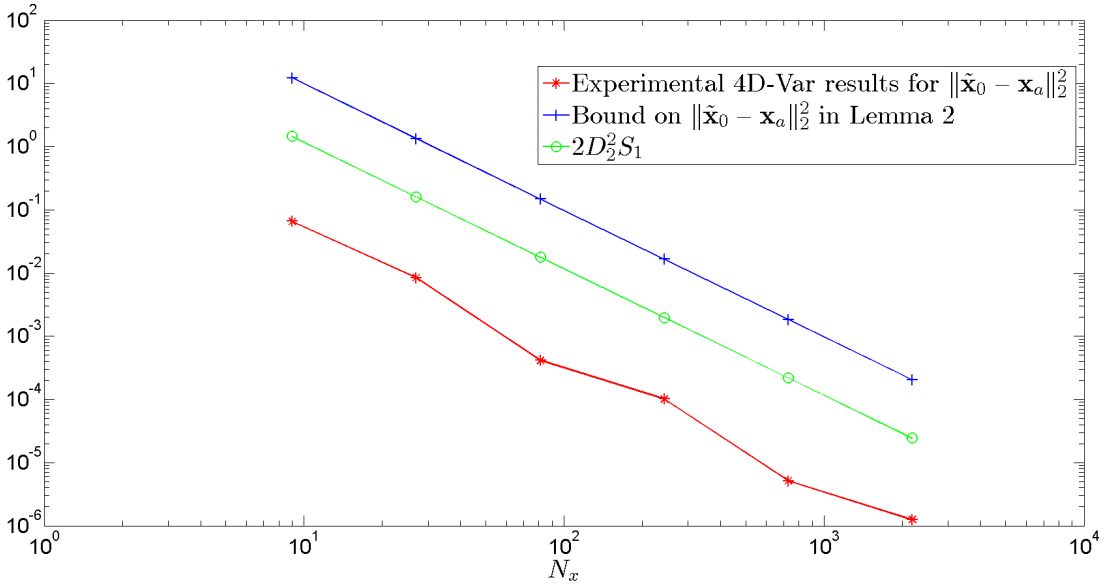


(a) For varying N_x, using fixed $L = 4$ ($\Delta t = \frac{1}{2N_x}$).

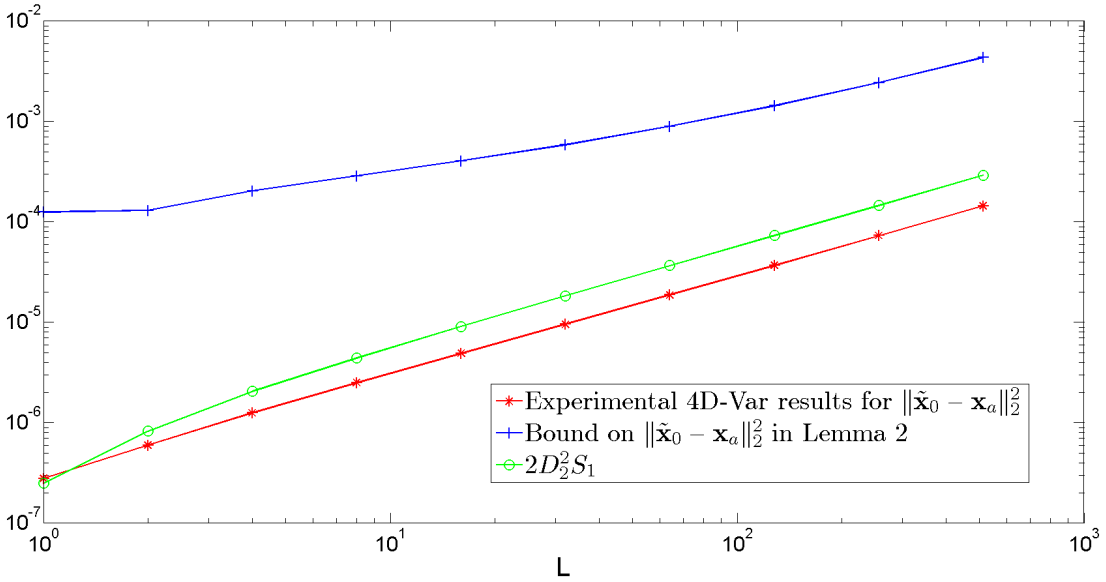


(b) For varying L, using fixed $N_x = 3^7$ ($\Delta t = \frac{1}{2\cdot 3^7}$).

Figure 4.1: *The square of the l_2-norm of the error in the analysis vector, as found through strong constraint 4D-Var data assimilation numerical experiments, is plotted alongside the bound in (4.19) for the same error in the analysis vector. The details of the numerical experiments are found in Section 4.3. The 1D square function initial condition in (4.28) is chosen chosen for use with the Upwind scheme, for demonstrating the effectiveness of the bound. The dominant summation $2D_2^2S_1$ of the bound in (4.19) is also plotted for comparison. When N_x is varied, the values of N_x are of the form $N_x = 3^\gamma$ where $\gamma = 2, \ldots, 7$. When L is varied, the values of L are of the form $L = 2^\delta$ where $\delta = 0, \ldots, 9$. The CFL number remained fixed with $h = 0.5$ and $\mu = 1$. The results are plotted using logarithmic scales to demonstrate the order of convergence.*



(a) For varying N_x, using fixed $L = 4$ ($\Delta t = \frac{1}{2N_x}$).



(b) For varying L, using fixed $N_x = 3^7$ ($\Delta t = \frac{1}{2\cdot 3^7}$).

Figure 4.2: *The square of the l_2-norm of the error in the analysis vector, as found through strong constraint 4D-Var data assimilation numerical experiments, is plotted alongside the bound in (4.19) for the same error in the analysis vector. The details of the numerical experiments are found in Section 4.3. The triangular function initial condition in (4.30) is chosen for use with the Preissman Box scheme, demonstrating the effectiveness of the bound. The dominant summation $2D_2^2S_1$ of the bound in (4.19) is also plotted for comparison. When N_x is varied, the values of N_x are of the form $N_x = 3^\gamma$ where $\gamma = 2, \ldots, 7$. When L is varied, the values of L are of the form $L = 2^\delta$ where $\delta = 0, \ldots, 9$. The CFL number remained fixed with $h = 0.5$ and $\mu = 1$. The results are plotted using logarithmic scales to demonstrate the order of convergence.*

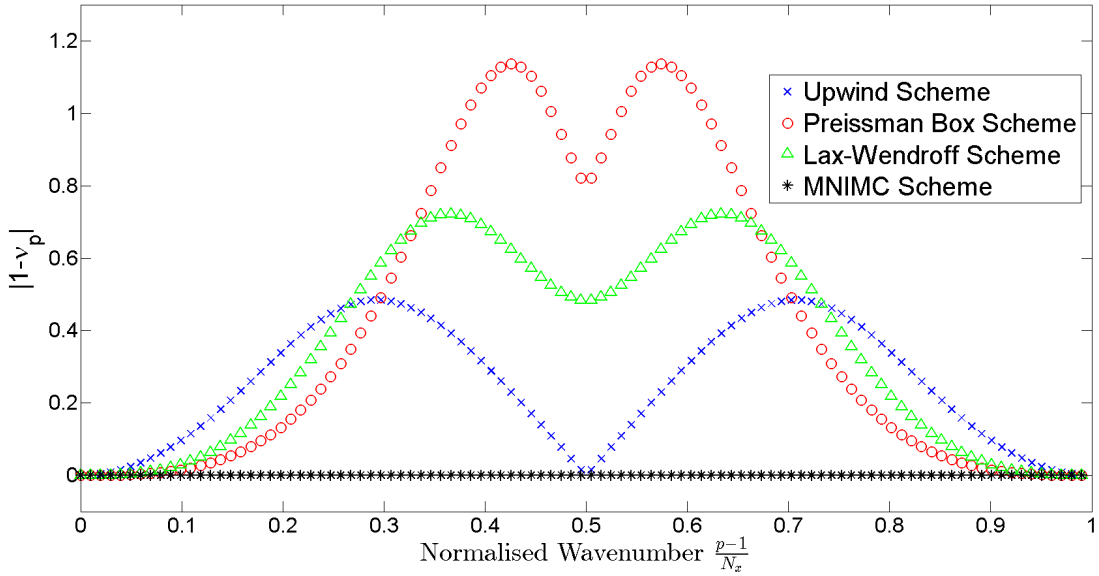
Figures 4.1 and 4.2 show that the bound described in (4.19) does seem to form a bound for the l_2-norm of the error in the analysis vector found through strong constraint 4D-Var numerical experiments, for the considered schemes, initial conditions and values of N_x and L. The bound does appear to be an order of magnitude larger than the numerical results, but it does appear to exhibit the same order of convergence. In order to investigate this apparent property, the bound in (4.19) is broken up into summations whose order of convergence can be investigated independently. This will allow us to identify which part of the bound, if any, behaves similarly to the error in the analysis vector.

As we wish to minimise the error in the analysis vector, we investigate the order of convergence of the error in the analysis vector and the summations to zero. The number of discretisation points (N_x) when considering full sets of observations and the number of sets of observations in the assimilation window (L) will be varied to investigate the order of convergence to zero with respect to these variables. We choose these variables as more discretisation points and more observations require greater computational resources, so it is important to understand if any gain in the accuracy of the analysis vector, is worth the extra expense. We also wish to understand how the regularity of the initial condition $u_0(x)$ and the numerically dissipative and dispersive properties of the finite difference schemes, affect the behaviour of the error and the bound. The next Section begins the process of identifying the order of convergence of the bound, by examining the coefficients $|1-\nu_p|$ and ξ_p, used in its construction.

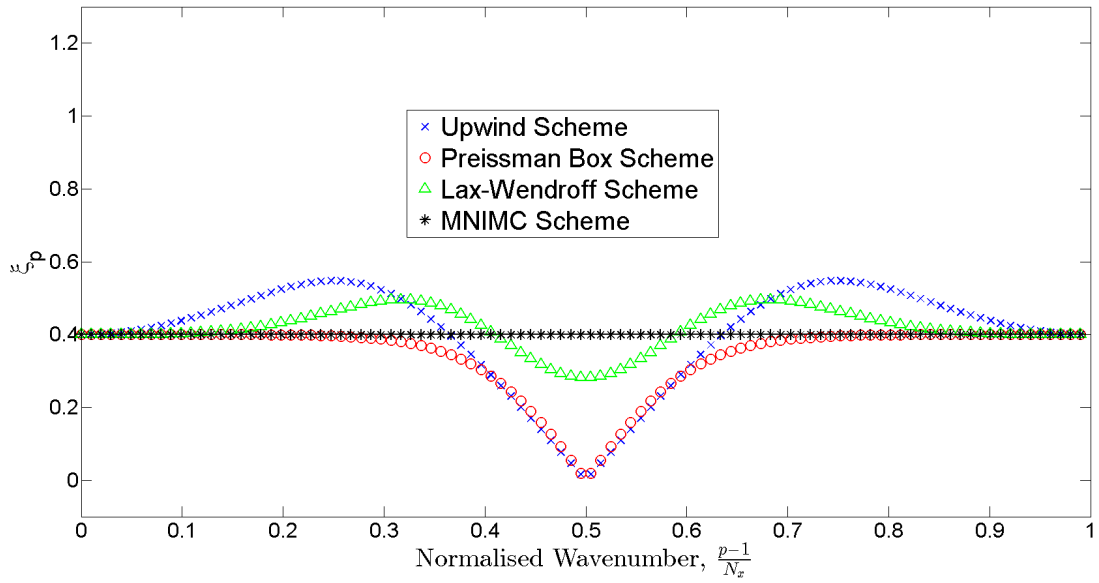
4.3.1 The order of convergence of $|1-\nu_p|$ and ξ_p

According to Section 4.3, the bound in (4.19) could potentially provide a good representation for the behaviour of the l_2-norm of the error in the analysis vector. The order of convergence of the bound with respect to either N_x or L, is in part determined by the order of convergence of the $|1-\nu_p|$ and ξ_p terms. The coefficient $|1-\nu_p|$ has a direct impact on the l_2-norm of the error in the analysis vector, whilst ξ_p is a consequence of applying a bound to the error in the analysis vector. The coefficients $|1-\nu_p|$ and ξ_p are both dependent on N_x and L; N_x determines the number of points, whilst L determines the shape of the plots in Figure 4.3.

The number of mesh points and observations used in NWP are typically $\mathcal{O}(10^7)$ [10] and $\mathcal{O}(10^5-10^6)$ [21] respectively. As a result, it is realistic to consider the order of convergence of the bound in (4.19) when L is small in comparison to N_x (ie: a small assimilation window). In the following Sections we will consider the order of convergence to zero of the coefficients $|1-\nu_p|$ and ξ_p with respect to N_x and L, whilst considering L to be small in comparison to N_x.



(a) $|1 - \nu_p|$



(b) ξ_p

Figure 4.3: *The values of* $|1 - \nu_p|$ *and* ξ_p *plotted against the corresponding normalised wavenumber ie:* $\frac{p-1}{N_x}$*, for* $p = 1, \ldots, N_x$*. The schemes considered are the Upwind, Preissman Box, Lax-Wendroff and MNIMC schemes for solving the 1D linear advection problem in (3.1), for* $h = 0.5$*,* $\mu = 1$*,* $N_x = 101$ *and* $L = 4$ *(*$\Delta t = \frac{1}{202}$*).*

The order of convergence with respect to N_x

Initially consider the order of convergence to zero of $|1 - \nu_p|$ and ξ_p with respect to N_x for fixed L. We will only consider $p = 2, \ldots, \frac{N_x+1}{2}$ for $|1 - \nu_p|$ and $p = 1, \ldots, \frac{N_x+1}{2}$ for ξ_p, as these are the only values of p which form a part of the bound in (4.19), and are not necessarily constant with respect to N_x and L. As N_x is increased, the shape of the plots in Figure 4.3 remain unchanged, but the number of points making up the

curves, increases. The effect of this is that for a fixed p, the corresponding value of $|1-\nu_p|$ and ξ_p are found to the left of their previous value, as N_x increases.

For example, let $N_x = N_x^{(1)}$ be odd and consider $|1-\nu_p|$ for some fixed p, $p = 2, \ldots, \frac{N_x+1}{2}$, which corresponds to $\frac{p-1}{N_x}$ along the bottom axis on Figure 4.3(a). Now increase N_x such that $N_x = N_x^{(2)} = 3N_x^{(1)}$. The effect of this is that two extra discretisation points are placed between each of the previous discretisation points along the bottom axis of Figure 4.3(a). Each one has a corresponding value of $|1-\nu_p|$. If we consider $|1-\nu_p|$ for the same fixed p, it corresponds to a new value along the bottom axis of Figure 4.3(a) as $\frac{p-1}{N_x^{(2)}} = \frac{p-1}{3N_x^{(1)}} < \frac{p-1}{N_x^{(1)}}$. This means that the new value of $|1-\nu_p|$ at $\frac{p-1}{N_x^{(2)}}$ is found to the left of its previous value at $\frac{p-1}{N_x^{(1)}}$. This gives $|1-\nu_p|$ an order of convergence with respect to N_x. The same is true for ξ_p. Increasing N_x results in $|1-\nu_p|$ and ξ_p moving left along the plots as we are considering fixed p, towards the left-most section of the plot where p is small in comparison to N_x.

The order of convergence of $|1-\nu_p|$ to zero with respect to N_x, for fixed p, was found numerically using fixed $L = 4$, $h = 0.5$ and $\mu = 1$ and increasing N_x in powers of three. Considering the order of convergence to zero with respect to N_x in the form of $\mathcal{O}(N_x^{\alpha_1})$ for $\alpha_1 \in \mathbb{R}$, α_1 was found to be less than or equal to zero for all $p = 2, \ldots, \frac{N_x+1}{2}$. As a result, $|1-\nu_p|$ is either decaying to zero or remaining constant as N_x increases. When considering fixed p, where p is initially close to $\frac{N_x+1}{2}$, α_1 varies until N_x is sufficiently large that p is small in comparison to N_x. This happened rapidly in these numerical experiments as N_x was increased in powers of three. When p was small in comparison to N_x ($\frac{p-1}{N_x} \ll 1$, $p \neq 1$), the order of convergence remained constant; $|1-\nu_p| = \mathcal{O}(N_x^{-2})$ for the Upwind scheme and $|1-\nu_p| = \mathcal{O}(N_x^{-3})$ for the Preissman Box and Lax-Wendroff schemes. These orders of convergence are determined by the gradient of the curves in Figure 4.3(a), for small p. This gradient is dictated by L, which determines the shape of the plots.

The order of convergence of ξ_p to zero with respect to N_x, for fixed p, was found numerically using fixed $L = 4$, $h = 0.5$ and $\mu = 1$ and increasing N_x in powers of three. We consider the order of convergence to zero of ξ_p with respect to N_x such that $\xi_p = \mathcal{O}(N_x^{\alpha_2})$ for $\alpha_2 \in \mathbb{R}$ and observe that α_2 is different for each p and appears to decay to zero as N_x increases. The value of α_2 is typically small, with order of magnitude $\mathcal{O}(10^{-1})$, taking both positive and negative values. The value of α_2 for mid-range values of p between 2 and $\frac{N_x+1}{2}$, was much larger with order of magnitude $\mathcal{O}(10^{1})$ for the Upwind and Preissman Box schemes. The value of α_2 decayed rapidly as N_x increased. A similar behaviour was seen for α_2 in the Lax-Wendroff scheme for p initially close to $\frac{N_x+1}{2}$. These orders of convergence are determined by the gradient of the curves in Figure 4.3(b), which are dictated by L.

The order of convergence with respect to L

We now consider the order of convergence to zero of $|1-\nu_p|$ and ξ_p with respect to L for fixed N_x. Again, we only consider $|1-\nu_p|$ for $p=2,\ldots,\frac{N_x+1}{2}$ and ξ_p for $p=1,\ldots,\frac{N_x+1}{2}$, for the same reasons as previously mentioned. As L increases, the shape of the curves in Figure 4.3 change whilst the number of points making up the curves, remains constant. Then for fixed p, the variables $|1-\nu_p|$ and ξ_p have an order of convergence with respect to L.

As L is increased, $|1-\nu_p|$ tends towards its limit,

$$|1-\nu_p| \to \begin{cases} 0, & \text{for } |\lambda_p|=1 \text{ and } \phi_p = 2\pi s,\ s\in\mathbb{Z}, \\ |\lambda_p|, & \text{for } |\lambda_p|<1, \text{ and } \phi_p = 2\pi s,\ s\in\mathbb{Z}, \\ 1, & \text{for } |\lambda_p|=1, \text{ and } \phi \neq 2\pi s,\ s\in\mathbb{Z}, \\ \left|1-\frac{(1-|\lambda_p|^2)(1-|\lambda_p|e^{i\phi_p})}{1+|\lambda_p|^2-2|\lambda_p|\cos(\phi_p)}\right|, & \text{for } |\lambda_p|<1, \text{ and } \phi_p \neq 2\pi s,\ s\in\mathbb{Z}. \end{cases} \tag{4.32}$$

When considering fixed N_x, the Upwind, Preissman Box and Lax-Wendroff schemes have the property that as $L\to\infty$, $|\lambda_2|$ is close to or equal to zero. Consequently as $L\to\infty$, $|1-\nu_2|$ is close to or equal to one for these three schemes. However, $|1-\nu_1|=0$ for all three schemes. This results in a steep gradient in the plot of $|1-\nu_p|$, for p small in comparison to N_x ($p\neq 1$), whose gradient increases as L increases. It is this gradient which provides the order of convergence with respect to N_x for small p, $p\neq 1$. Obviously the limit of $|1-\nu_p|$ as $L\to\infty$ is not always zero, however we are still interested in the l_2-norm of the error in the analysis vector behaves with respect to zero. Therefore we will continue to consider the order of convergence of $|1-\nu_p|$ to zero with respect to L.

The order of convergence of $|1-\nu_p|$ to zero with respect to L, for fixed p, was found numerically using fixed $N_x=3^7$, $h=0.5$ and $\mu=1$ and increasing L in powers of two. Considering the order of convergence to zero with respect to L in the form $|1-\nu_p|=\mathcal{O}(L^{\beta_1})$ for $\beta_1\in\mathbb{R}$, β_1 was found to be positive for all $p=2,\ldots,\frac{N_x+1}{2}$ and at most $|1-\nu_p|=\mathcal{O}(L)$ for all three schemes. This order of convergence was achieved for p small in comparison to N_x, $p\neq 1$. These results show that increasing the value of L causes the value of $|1-\nu_p|$ to diverge from zero for $p\neq 1$ as demonstrated by the limit in (4.32).

The value of β_1 was found to be extremely small for p close to $\frac{N_x+1}{2}$ for each scheme. Section 3.10 mentioned that when p is close to $\frac{N_x+1}{2}$, ν_p approaches its limit for $L\to\infty$ for a relatively small value of L, for the considered schemes. Consequently, the order of convergence with respect to L for p close to $\frac{N_x+1}{2}$, would be quite small. The similar behaviour in the orders of convergence with respect to L for each scheme ($p\neq 1$), is not surprising as the dependence of $|1-\nu_p|$ on L is similar for each scheme, unlike their dependence on N_x.

In the case of the variable ξ_p as defined in equation (4.20), it is not appropriate to consider ξ_p as $L\to\infty$. This variable is defined through the creation of the bound on $\boldsymbol{\rho}_L$.

The form of $\boldsymbol{\rho}_L$ defined in (3.72) of Lemma 3.13, relies upon L being finite. Therefore, ξ_p only exists in this form when L is finite so cannot be considered as $L \to \infty$. As $L \to \infty$, we can use the result of Lemma 3.13 and re-write $\mathbf{s}_L$ in equation (3.74) to create,

$$\boldsymbol{\rho}_\infty = \lim_{L\to\infty} V \left[\sum_{k=0}^{L} (\Lambda^*\Lambda)^k\right]^{-1} \left[\sum_{l=0}^{L} (\Lambda^*\tilde{\Lambda})^{lb}\right] \left[\sum_{y=1}^{b-1} (\Lambda^*)^j V^* \mathbf{r}_j\right], \tag{4.33}$$

where $\sum_{y=1}^{b-1}(\Lambda^*)^j V^*\mathbf{r}_j$ is defined to be zero when $b = 1$. We can now find $|\mathcal{F}_p(\boldsymbol{\rho}_\infty)|$ in the same way as was done for finite L in (4.25),

$$\begin{aligned} |\mathcal{F}_p(\boldsymbol{\rho}_\infty)| &\leq \lim_{L\to\infty} \frac{\left|\sum_{l=0}^{L}(|\lambda_p|^b e^{ib\phi_p})^l\right| \left(\sum_{j=1}^{b-1} |\lambda_p|^j |\mathcal{F}_p(\mathbf{r}_j)|\right)}{\sum_{k=0}^{L} |\lambda_p|^{2k}}, \\ &\leq \frac{2\sqrt{N_x} D_3 \lim_{L\to\infty} \xi_p}{N_x^{r+1}}, \end{aligned} \tag{4.34}$$

by (4.26) for all $p = 1, \ldots, N_x$. Now consider,

$$\lim_{L\to\infty} \xi_p := \lim_{L\to\infty} \frac{\left|\sum_{l=0}^{L}(|\lambda_p|^b e^{ib\phi_p})^l\right| \left(\sum_{j=1}^{b-1} |\lambda_p|^j\right)}{\sum_{k=0}^{L} |\lambda_p|^{2k}} \tag{4.35}$$

$$= \begin{cases} 0, & \text{for } |\lambda_p| = 1, \\ \frac{|\lambda_p|(1-|\lambda_p|^{b-1})(1+|\lambda_p|)}{\sqrt{1+|\lambda_p|^{2b}-2|\lambda_p|^b\cos(b\phi_p)}}, & \text{for } |\lambda_p| < 1, \end{cases} \tag{4.36}$$

for $p = 1, \ldots, N_x$. Here we can see that it is the dissipative properties of the finite difference scheme that determine the form of ξ_p as $L \to \infty$.

When the scheme is non-dissipative, such as for the Preissman Box and the MNIMC schemes, $\lim_{L\to\infty} \xi_p = 0$. This results in the contribution from aliasing errors to the bound in (4.19), decaying to zero. In Section 3.10.5, we discussed that $\lim_{L\to\infty} \mathbf{x}_a = \lim_{L\to\infty} \boldsymbol{\rho}_L$ for the Preissman Box scheme. However, we were not able to say how $\boldsymbol{\rho}_L$ behaved as $L \to \infty$ because we were not able to ascertain how $\mathbf{r}_j$ behaved at this limit for $j = 1, \ldots, b-1$. Now we know that $\mathbf{x}_a \to \mathbf{0}$ as $L \to \infty$ for a numerically non-dissipative finite difference scheme with respect to the resolvable wavenumber components of the numerical solution, as the upper bound on the Fourier coefficients of $\boldsymbol{\rho}_\infty$ decays to zero as $L \to \infty$ in Equation (4.34). In this instance destructive interference is affecting all wavenumber components of the solution, attenuating them all to zero as L increases, resulting in a loss of information from the analysis vector.

In the case of the MNIMC scheme, we see that increasing the number of sets of observations in the assimilation window, decreases the aliasing error in the analysis vector, as the upper bound on the Fourier coefficients of the error decay to zero. Since this is the only error present in the analysis vector for this scheme ($h \neq 1$), as more sets of observations are provided in the assimilation window, the analysis vector will become closer to the discrete sample of the true initial condition we wish to recover.

In this instance, the scheme behaves as we had intuitively expected all of our schemes to behave, with respect to extra sets of observations in the assimilation window.

If we now consider a dissipative scheme with respect to the resolvable wavenumber components of the numerical solution, we can see that as $L \to \infty$, ξ_p does not tend towards zero. Therefore the bound in (4.19) may not tend towards zero as $L \to \infty$ in this instance. This means that no matter how many sets of observations we use in the assimilation window, the error in the analysis vector may never be zero.

Even though $\boldsymbol{\rho}_L$ cannot be considered directly as $L \to \infty$, we can still examine its order of convergence to zero, for finite values of L used in numerical experiments. The order of convergence of ξ_p to zero with respect to L, for fixed p, was found numerically using fixed $N_x = 3^7$, $h = 0.5$ and $\mu = 1$ and considered in the form $\xi_p = \mathcal{O}(L^{\beta_2})$ for $\beta_2 \in \mathbb{R}$. For small p ($p \neq 1$) β_2 was initially $\mathcal{O}(10^{-1})$, but decayed rapidly to zero as L increased. Examining β_2 for larger values of p close to $\frac{N_x+1}{2}$, we found the rate of decay of β_2 occurred much faster.

In the following Section we attempt to use asymptotic expansions to explicitly represent the dependence of $|1 - \nu_p|$ and ξ_p with respect to the number of discretisation points N_x when considering full sets of observations and the number of sets of observations in the assimilation window L. As we wish to use the bound in (4.19) to characterise the behaviour of the error in the analysis vector, it would be nice to have an analytical form for $|1 - \nu_p|$ and ξ_p to explicitly demonstrate their dependence and hence the dependence of the bound, on N_x and L. Otherwise, the only way to characterise the behaviour of the bound is to numerically generate it.

4.3.2 Asymptotic expansions of $|1 - \nu_p|$ and ξ_p

The numerical order of convergence for $|1 - \nu_p|$ to zero with respect to both N_x and L, can be explained through its asymptotic expansion as $N_x \to \infty$, for fixed $p = 2, \ldots, \frac{N_x+1}{2}$. We will demonstrate this using the Upwind scheme when $h = 0.5$ where,

$$|\lambda_p| = \cos\left(\frac{\pi(p-1)}{N_x}\right),$$

for $p = 2, \ldots, \frac{N_x+1}{2}$. Assuming $L > 0$, let $z = \frac{p-1}{N_x}$. If we consider a fixed p, as $N_x \to \infty$, $z \to 0$, so we consider z as a continuous variable. Taylor expanding $|1 - \nu(z)|$ about $z = 0$, results in,

$$|1 - \nu(z)| = \frac{\pi^2 L}{4} z^2 + \mathcal{O}(z^4), \quad \text{for } 0 < z < \frac{1}{2},$$

as the 4th derivative of $|1 - \nu(z)|$ with respect to z is continuous over $0 < z < \frac{1}{2}$, so is bounded on this domain. Considering $z = \frac{p-1}{N_x}$ for $1 < p < \frac{N_x+1}{2}$ as $N_x \to \infty$, we

obtain,

$$|1-\nu_p| = \frac{\pi^2 L}{4}\left(\frac{p-1}{N_x}\right)^2 + \mathcal{O}\left(\left[\frac{p-1}{N_x}\right]^4\right). \tag{4.37}$$

Therefore,

$$|1-\nu_p| \sim \frac{\pi^2 L}{4}\left(\frac{p-1}{N_x}\right)^2, \text{ as } N_x \to \infty. \tag{4.38}$$

We will trial the use of $|1-\nu_p| \approx \frac{\pi^2 L}{4}\left(\frac{p-1}{N_x}\right)^2 = \mathcal{O}\left(L\left[\frac{p-1}{N_x}\right]^2\right)$ for $p = 2, \ldots, \frac{N_x+1}{2}$.

This expansion indicates that $|1-\nu_p|$ has orders of convergence $\mathcal{O}(N_x^{-2})$ and $\mathcal{O}(L)$ for the Upwind scheme when p is small ($p \neq 1$). These match the numerical orders of convergence found for $|1-\nu_p|$ to zero with respect to N_x and L when p is small ($p \neq 1$), for the Upwind scheme. Similar Taylor expansions can be constructed for $|1-\nu_p|$ for the Preissman Box and Lax-Wendroff schemes. However the form of ν_p in equation (3.76) for numerically dissipative schemes with respect to the resolvable wavenumber components of the numerical solution, does not lend itself to differentiation when considering $|1-\nu(z)|$. Several applications of L'Hôpital's rule is required to evaluate each derivative at $z = 0$.

In Section 4.3.3 we will try to represent the order of convergence of the bound on the l_2-norm of the error in the analysis vector analytically, with respect to both N_x and L. As we only have a Taylor expansion for $|1-\nu_p|$ for the Upwind scheme, we will only perform this analysis for the Upwind scheme. Once a Taylor expansion has been achieved for the Preissman Box and Lax-Wendroff schemes, the analysis in that Section can be performed in the same way. The right-hand side of (4.37) is zero when $p = 1$, which is identical to $|1-\nu_1|$, as previously obtained. Retaining the variable p on the right-hand side allows the order of convergence to zero to change with p, representing that we only achieved an order of convergence of $\mathcal{O}(LN_x^{-2})$ numerically, when p was small ($p \neq 1$) for the Upwind scheme.

As for $|1-\nu_p|$, the numerical orders of convergence of ξ_p to zero with respect to N_x and L for small p, can be explained through the asymptotic expansion of ξ_p. We consider a Taylor expansion of ξ_p as $N_x \to \infty$, for fixed $p = 2, \ldots, \frac{N_x+1}{2}$. Assuming $L > 0$, we again let $z = \frac{p-1}{N_x}$. Then using the same reasoning as for the Taylor expansion of $|1-\nu(z)|$ about $z = 0$ in (4.37), we Taylor expand $\xi(z)$ about $z = 0$ resulting in,

$$\xi(z) = \frac{L(b-1)+[L]_b}{(L+1)b} + \mathcal{O}(z^2), \text{ for } 0 < z < \frac{1}{2},$$

for $b \in \mathbb{N}\setminus\{1\}$, as the 2nd derivative of the function $\xi(z)$ is continuous over $0 < z < \frac{1}{2}$, so is bounded on this domain. If $b = 1$, then $\xi(z) = 0$. Considering $z = \frac{p-1}{N_x} < \frac{1}{2}$ as $N_x \to \infty$, for $b \in \mathbb{N}\setminus\{1\}$ we obtain,

$$\xi_p = \frac{L(b-1)+[L]_b}{(L+1)b} + \mathcal{O}\left(\left[\frac{p-1}{N_x}\right]^2\right). \tag{4.39}$$

Therefore,

$$\xi_p \sim \frac{L(b-1)+[L]_b}{(L+1)b}, \text{ as } N_x \to \infty. \tag{4.40}$$

We trial the use of $\xi_p \approx \frac{L(b-1)+[L]_b}{(L+1)b} = \mathcal{O}(1)$ for $p = 1, \ldots, \frac{N_x+1}{2}$. This does not completely represent the orders of convergence found numerically for ξ_p with respect to N_x and L for $p = 1, \ldots, \frac{N_x+1}{2}$. The variable ξ_p has a small order of convergence for most p, which decays to zero as either N_x or L increase. Therefore, $\xi_p = \mathcal{O}(1)$ maybe sufficient analytically for large enough N_x or L, but we will continue with the knowledge that we have not completely characterised the behaviour of ξ_p for $p = 1, \ldots, \frac{N_x+1}{2}$.

4.3.3 Analysis of the summations comprising the bound on the error in the analysis vector

The bound on the l_2-norm of the error in the analysis vector in (4.19), can be re-written in terms of the sum of individual summations. Schemes that are numerically dissipative and/or dispersive with respect to the resolvable wavenumbers of the solution, generally do not have the property that $|1-\nu_p| = 0$ for $p = 2, \ldots, N_x$. When considering the Upwind, Preissman Box and Lax-Wendroff schemes $|1-\nu_1| = 0$, so re-writing the bound in this way for these schemes, results in the bound being comprised of the sum of six distinct summations,

$$\|\tilde{\mathbf{x}}_0 - \mathbf{x}_a\|_2^2 \leq 2D_2^2 S_1 + 4D_2 D_3 S_2 + 2D_3^2 S_3 + 4D_3^2 S_4 + 8D_3^2 S_5 + 8D_2 D_3 S_6, \tag{4.41}$$

where,

$$\begin{aligned}
S_1 &= N_x \textstyle\sum_{p=2}^{\frac{N_x+1}{2}} \frac{|1-\nu_p|^2}{(p-1)^{2(r+1)}}, & S_2 &= \frac{1}{N_x^r} \textstyle\sum_{p=2}^{\frac{N_x+1}{2}} \frac{|1-\nu_p|^2}{(p-1)^{r+1}}, \\
S_3 &= \frac{1}{N_x^{2r+1}} \textstyle\sum_{p=2}^{\frac{N_x+1}{2}} |1-\nu_p|^2, & S_4 &= \frac{1}{N_x^{2r+1}} \left(\xi_1^2 + 2 \textstyle\sum_{p=2}^{\frac{N_x+1}{2}} \xi_p^2 \right), \\
S_5 &= \frac{1}{N_x^{2r+1}} \textstyle\sum_{p=2}^{\frac{N_x+1}{2}} |1-\nu_p| \xi_p, & S_6 &= \frac{1}{N_x^r} \textstyle\sum_{p=2}^{\frac{N_x+1}{2}} \frac{|1-\nu_p| \xi_p}{(p-1)^{r+1}}.
\end{aligned} \tag{4.42}$$

Each of these summations is positive in value and dependent on the regularity (r) of the true initial condition $u_0(x)$ over $(0,1)$. The coefficients D_2 and D_3 are not considered as a part of the summations as they are constant with respect to N_x and L. The coefficient D_1 does not appear in Equation (4.41) as it corresponds to ν_1, which does not appear in the Equation as $|1-\nu_1| = 0$ for the considered schemes.

When considering a scheme which is numerically non-dissipative and non-dispersive with respect to the resolvable wavenumber components of the solution, such as for the MNIMC scheme, $|1-\nu_p| = 0$ for all $p = 1, \ldots, N_x$. As a result, only one summation of (4.19) is non-zero. This gives,

$$\|\tilde{\mathbf{x}}_0 - \mathbf{x}_a\|_2^2 \leq 4D_3^2 S_4. \tag{4.43}$$

Each summation has an order of convergence to zero with respect to N_x and L, which influences the overall order of convergence for the bound. In the following Sections, we identify the order of convergence of each summation in (4.42) to zero, with respect to N_x and L, numerically. This is done for the Upwind, Preissman Box, Lax-Wendroff and MNIMC schemes, using initial conditions with different regularities. This will allow us to identify which summations are dominant, determining the behaviour of the bound on the l_2-norm of the error in the analysis vector. The dominant summations can then be analysed to determine the behaviour of the bound. We also compare the numerical orders of convergence to zero with the analytical orders of convergence for the Upwind scheme, found using (4.38) and (4.40). It is important to note that in order to identify the numerical orders of convergence with respect to a given variable, the variable needs to be sufficiently large such that the numerical results converge to the order of converge we are looking for. In the following Tables, the values for the numerical orders of convergence to zero were identified using the largest values of N_x and L considered in the experiments. The orders of convergence with respect to N_x had converged by this point, but not the orders of convergence with respect to L.

The order of convergence of S_1

$$S_1 = N_x \sum_{p=2}^{\frac{N_x+1}{2}} \frac{|1-\nu_p|^2}{(p-1)^{2(r+1)}}$$

This summation is composed from the amplification factors and the bound on the Fourier coefficients of $u_0(x)$, giving it an explicit dependence on the regularity (r) of $u_0(x)$ over $(0,1)$. The orders of convergence to zero for this summation, with respect to both N_x and L found numerically, are given in Table 4.1.

Table 4.1 shows that the order of convergence of S_1 to zero with respect to N_x, for an initial condition with $r = 0$, is $\mathcal{O}(N_x^0)$ for all schemes. This indicates that the summation remains constant with respect to N_x when $r = 0$. For higher regularity initial conditions, the order of convergence is less than zero for each scheme. This indicates that S_1 is decaying towards zero as N_x increases.

As the regularity is increased, the order of convergence to zero with respect to N_x is initially $\mathcal{O}(N_x^{-2r})$ for each scheme. However, once a critical regularity is achieved the order of convergence saturates; $\mathcal{O}(N_x^{-3})$ for the Upwind scheme when $r \geq 2$ and $\mathcal{O}(N_x^{-5})$ for the Preissman Box and Lax-Wendroff scheme when $r \geq 3$.

Table 4.1 also shows that the order of convergence of S_1 to zero, with respect to L, is positive for all values of r. This indicates that S_1 increases as the length of the assimilation window is increased. The order of convergence also increases with the value of r associated with the initial condition, until a critical value is reached, where the order of convergence saturates at $\mathcal{O}(L^2)$. The regularity at which the saturation point is reached is $r = 2$ for all schemes, since they all have a similar dependence on L.

Representing $|1-\nu_p|$ using (4.38) for the Upwind scheme, Appendix B.1.3 derives

the analytical order of convergence to zero for S_1, for the Upwind scheme in Equation (4.44). These analytical orders of convergence to zero match the numerical results in Table 4.1 for the Upwind scheme with respect to N_x and the saturation order of convergence with respect to L. The failure to analytically capture the behaviour with respect to L for small r will be discussed in Section 4.3.4.

r	α			β		
	Upwind	Preissman Box	Lax-Wendroff	Upwind	Preissman Box	Lax-Wendroff
0	5.1800×10^{-12}	-1.7666×10^{-6}	-2.4129×10^{-9}	5.8528×10^{-1}	3.8727×10^{-1}	4.1700×10^{-1}
1	-1.9984	-2.0000	-2.0000	1.5054	9.9622×10^{-1}	1.0236
2	-2.9977	-3.9992	-4.0180	1.9957	1.6588	1.6731
3	-3.0000	-4.9999	-5.0196	2.0000	1.9991	1.9955
4	-3.0000	-5.0000	-5.0000	2.0000	2.0000	2.0000
5	-3.0000	-5.0000	-5.0000	2.0000	2.0000	2.0000
6	-3.0000	-5.0000	-5.0000	2.0000	2.0000	2.0000
7	-3.0000	-5.0000	-5.0000	2.0000	2.0000	2.0000
$r \gg 1$	-3.0000	-5.0000	-5.0000	2.0000	2.0000	2.0000

Table 4.1: *The numerical orders of convergence to zero, with respect to N_x and L, for $S_1 = \mathcal{O}(N_x^\alpha L^\beta)$, using the Upwind, Preissman Box and Lax-Wendroff schemes, given to 4dp (decimal places), for $h = 0.5$ and $\mu = 1$. The results for N_x were identified using fixed $L = 4$ ($\Delta t = \frac{1}{2N_x}$), considering N_x in the form $N_x = 3^\gamma$, where $\gamma = 2, \ldots, 7$ for the Upwind and Preissman Box schemes and $\gamma = 2, \ldots, 12$ for the Lax-Wendroff scheme. The results for L were identified using fixed $N_x = 3^7$ ($\Delta t = \frac{1}{2 \cdot 3^7}$), considering L in the form $L = 2^\delta$, where $\delta = 0, \ldots, 9$. The results for $r \gg 1$ were identified using (4.52). The results displayed here are the orders of convergence for the largest values of N_x and L considered, respectively. The full table of orders of convergence can be found in Appendix B.1.*

$$S_1 = \begin{cases} \mathcal{O}(L^2 N_x^{-2r}), & \text{for } r = 0, 1, \\ \mathcal{O}(L^2 N_x^{-3}), & \text{for } r \geq 2. \end{cases} \tag{4.44}$$

The order of convergence of S_2

$$S_2 = \frac{1}{N_x^r} \sum_{p=2}^{\frac{N_x+1}{2}} \frac{|1-\nu_p|^2}{(p-1)^{r+1}}$$

This summation is composed from the amplification factors and the bounds on the Fourier coefficients of $u_0(x)$ and the error in the 1D DFT of $u_0(x)$, giving it an explicit dependence on the regularity (r) of $u_0(x)$ over $(0,1)$ and the number of discretisation points (N_x) when considering full sets of observations. The numerical orders of convergence to zero for this summation, with respect to both N_x and L, are given in Table 4.2.

Table 4.2 shows that the order of convergence of S_2 to zero with respect to N_x, for all regularity initial conditions $u_0(x)$, is less than or equal to zero. When $r = 0$, the error is $\mathcal{O}(N_x^0)$, indicating that the error does not decay as N_x is increased. Higher regularity initial conditions all have a negative order of convergence showing that the error decays to zero as N_x increases. S_2 appears to have an order of convergence $\mathcal{O}(N_x^{-2r})$ until $r = 4$ for the Upwind scheme and $r = 6$ for the Preissman Box and Lax-Wendroff schemes. The results for the Upwind scheme can be explained by examining the orders of convergence for S_2 analytically.

Representing $|1 - \nu_p|$ using (4.38) for the Upwind scheme, Appendix B.2.3 derives Equation (4.45). These analytical results match the numerical results in Table 4.2 for the order of convergence of S_2 to zero, for the Upwind scheme, with respect to N_x. This shows that when $r = 4$ the order of convergence with respect to N_x has a $\log(N_x)$ factor associated with it, causing the small variation in Table 4.2 at $r = 4$ for the Upwind scheme. A similar factor may be affecting the Preissman Box and Lax-Wendroff schemes when $r = 6$. This will be discussed in Section 4.3.4.

The numerical orders of convergence with respect to L in Table 4.2, show that S_2 increases as L increases. The orders of convergence also appear to show a saturation point of $\mathcal{O}(L^2)$ once a sufficient regularity has been reached for each scheme. They also show that the orders of convergence with respect to L for each scheme are similar. However as with S_1, the analytical order of convergence for the Upwind scheme, only captures the saturated order of convergence rather than the changes with regularity we see in Table 4.2.

r	α			β		
	Upwind	Preissman Box	Lax-Wendroff	Upwind	Preissman Box	Lax-Wendroff
0	-9.8321×10^{-12}	-1.9697×10^{-6}	-3.5133×10^{-11}	2.4804×10^{-1}	1.6992×10^{-1}	1.9195×10^{-1}
1	-2.0000	-2.0000	-2.0000	5.8528×10^{-1}	3.8727×10^{-1}	4.1700×10^{-1}
2	-4.0000	-4.0000	-4.0000	1.0274	6.7655×10^{-1}	7.0649×10^{-1}
3	-5.9984	-6.0000	-6.0000	1.5054	9.9622×10^{-1}	1.0236
4	-7.8407	-8.0000	-8.0000	1.8998	1.3256	1.3492
5	-8.9977	-9.9993	-10.0180	1.9957	1.6588	1.6732
6	-10.0000	-11.8630	-12.0201	1.9999	1.9426	1.9359
7	-11.0000	-12.9999	-13.0196	2.0000	1.9991	1.9955

Table 4.2: *The numerical orders of convergence to zero, with respect to N_x and L, for $S_2 = \mathcal{O}(N_x^{\alpha} L^{\beta})$, using the Upwind, Preissman Box and Lax-Wendroff schemes, given to 4dp, for $h = 0.5$ and $\mu = 1$. The results for N_x were identified using fixed $L = 4$ ($\Delta t = \frac{1}{2N_x}$), considering N_x in the form $N_x = 3^{\gamma}$, where $\gamma = 2, \ldots, 7$ for the Upwind and Preissman Box schemes and $\gamma = 2, \ldots, 12$ for the Lax-Wendroff scheme. The results for L were identified using fixed $N_x = 3^7$ ($\Delta t = \frac{1}{2 \cdot 3^7}$), considering L in the form $L = 2^{\delta}$, where $\delta = 0, \ldots, 9$. The results displayed here are the orders of convergence for the largest values of N_x and L considered, respectively. The full table of orders of convergence can be found in Appendix B.2.*

$$S_2 = \begin{cases} \mathcal{O}(L^2 N_x^{-2r}), & \text{for } r = 0, 1, 2, 3, \\ \mathcal{O}(L^2 N_x^{-2r} \log(N_x)), & \text{for } r = 4 \\ \mathcal{O}(L^2 N_x^{-4-r}), & \text{for } r \geq 5. \end{cases} \tag{4.45}$$

The order of convergence of S_3

$$S_3 = \frac{1}{N_x^{2r+1}} \sum_{p=2}^{\frac{N_x+1}{2}} |1-\nu_p|^2$$

This summation is composed from the amplification factors and the bound on the error in the 1D DFT of $u_0(x)$, giving it an explicit dependence on the regularity (r) of $u_0(x)$ over $(0,1)$ and the number of discretisation points (N_x) when considering full sets of observations. The numerical orders of convergence to zero for this summation, with respect to both N_x and L, are given in Table 4.3.

The numerical orders of convergence with respect to N_x for each scheme, are identical for each value of r. This is due to the summation portion of S_3, being independent of r. The numerical results give that S_3 has order of convergence $\mathcal{O}(N_x^{-2r})$.

The numerical order of convergence to zero with respect to L in Table 4.3, is similar in order of magnitude for each scheme and is constant with respect to regularity. The orders of convergence for L in this table have not converged by this point and are continuing to change as L is increased. This can be seen in the full table of results in Appendix B.3.

Representing $|1-\nu_p|$ using (4.38), Appendix B.3.3 derives the analytical order of convergence to zero of S_3 for the Upwind scheme,

$$S_3 = \mathcal{O}(L^2 N_x^{-2r}). \tag{4.46}$$

This analytical result matches the numerical orders of convergence with respect to N_x in Table 4.3 for the Upwind scheme. However, we do not capture the behaviour of S_3 with respect to L.

The orders of convergence to zero with respect to L for summations S_1 and S_2, vary with regularity of the initial condition. Summation S_3 does not have this property. It differs from these summations by not possessing a $(p-1)^{-(r+1)}$ term. Therefore it must be the interaction of $(p-1)^{-(r+1)}$ with $|1-\nu_p|$ and the summation which changes the order of convergence with respect to L in S_1 and S_2, for varying regularity true initial conditions. As a result, there is some dependence of the $\frac{p-1}{N_x}$ terms on L in $|1-\nu_p|$, that the asymptotic expansion of $|1-\nu_p|$ does not capture.

r	α			β		
	Upwind	Preissman Box	Lax-Wendroff	Upwind	Preissman Box	Lax-Wendroff
0	4.0423×10^{-16}	-2.0610×10^{-6}	-1.9489×10^{-6}	7.2171×10^{-2}	5.3321×10^{-2}	6.2698×10^{-2}
1	-2.0000	-2.0000	-2.0000	7.2171×10^{-2}	5.3321×10^{-2}	6.2698×10^{-2}
2	-4.0000	-4.0000	-4.0000	7.2171×10^{-2}	5.3321×10^{-2}	6.2698×10^{-2}
3	-6.0000	-6.0000	-6.0000	7.2171×10^{-2}	5.3321×10^{-2}	6.2698×10^{-2}
4	-8.0000	-8.0000	-8.0000	7.2171×10^{-2}	5.3321×10^{-2}	6.2698×10^{-2}
5	-10.0000	-10.0000	-10.0000	7.2171×10^{-2}	5.3321×10^{-2}	6.2698×10^{-2}
6	-12.0000	-12.0000	-12.0000	7.2171×10^{-2}	5.3321×10^{-2}	6.2698×10^{-2}
7	-14.0000	-14.0000	-14.0000	7.2171×10^{-2}	5.3321×10^{-2}	6.2698×10^{-2}

Table 4.3: *The numerical orders of convergence to zero, with respect to N_x and L, for $S_3 = \mathcal{O}(N_x^{\alpha} L^{\beta})$, using the Upwind, Preissman Box and Lax-Wendroff schemes, given to 4dp, for $h = 0.5$ and $\mu = 1$. The results for N_x were identified using fixed $L = 4$ ($\Delta t = \frac{1}{2N_x}$), considering N_x in the form $N_x = 3^{\gamma}$, where $\gamma = 2, \ldots, 7$. The results for L were identified using fixed $N_x = 3^7$ ($\Delta t = \frac{1}{2 \cdot 3^7}$), considering L in the form $L = 2^{\delta}$, where $\delta = 0, \ldots, 9$. The results displayed here are the orders of convergence for the largest values of N_x and L considered, respectively. The full table of orders of convergence can be found in Appendix B.3.*

The order of convergence of S_4

$$S_4 = \frac{1}{N_x^{2r+1}} \left(\xi_1^2 + 2 \sum_{p=2}^{\frac{N_x+1}{2}} \xi_p^2 \right)$$

This summation is composed from the bounds on the error in the 1D DFT of $u_0(x)$ and the aliasing error due to the MNIMC scheme, giving it an explicit dependence on the regularity (r) of $u_0(x)$ over $(0, 1)$ and the number of discretisation points (N_x) when considering full sets of observations. The numerical orders of convergence to zero for this summation, with respect to both N_x and L, are given in Table 4.4 for the Upwind, Preissman Box and Lax-Wendroff schemes.

These numerical results give that S_4 has a numerical order of convergence to zero with respect to N_x of $\mathcal{O}(N_x^{-2r})$. Hence when $r = 0$, S_4 does not decay to zero, whilst for higher regularity initial conditions, S_4 does decay to zero as N_x increases.

Representing ξ_p using (4.40), Appendix B.4.3 derives the analytical order of convergence of S_4 to zero for the Upwind scheme,

$$S_4 = \mathcal{O}(N_x^{-2r}). \tag{4.47}$$

This analytical order of convergence to zero matches the numerical results in Table 4.4 for the Upwind scheme with respect to N_x.

The numerical order of convergence to zero with respect to L in Table 4.4, is constant with respect to regularity. As with S_3, the orders of convergence with respect to L in this Table have yet to finish converging as L is increased, but it does not appear that they are converging to any particular rate. The numerical orders of convergence with respect to L are not zero for the Upwind scheme, so the analytical order of convergence to zero with respect to L is not capturing the behaviour of S_4 with respect to L.

r	α			β		
	Upwind	Preissman Box	Lax-Wendroff	Upwind	Preissman Box	Lax-Wendroff
0	-6.3666×10^{-15}	-2.0211×10^{-16}	-1.9397×10^{-11}	4.4885×10^{-2}	-3.2174×10^{-1}	-8.8338×10^{-2}
1	-2.0000	-2.0000	-2.0000	4.4885×10^{-2}	-3.2174×10^{-1}	-8.8338×10^{-2}
2	-4.0000	-4.0000	-4.0000	4.4885×10^{-2}	-3.2174×10^{-1}	-8.8338×10^{-2}
3	-6.0000	-6.0000	-6.0000	4.4885×10^{-2}	-3.2174×10^{-1}	-8.8338×10^{-2}
4	-8.0000	-8.0000	-8.0000	4.4885×10^{-2}	-3.2174×10^{-1}	-8.8338×10^{-2}
5	-10.0000	-10.0000	-10.0000	4.4885×10^{-2}	-3.2174×10^{-1}	-8.8338×10^{-2}
6	-12.0000	-12.0000	-12.0000	4.4885×10^{-2}	-3.2174×10^{-1}	-8.8338×10^{-2}
7	-14.0000	-14.0000	-14.0000	4.4885×10^{-2}	-3.2174×10^{-1}	-8.8338×10^{-2}

Table 4.4: *The numerical orders of convergence to zero, with respect to N_x and L, for $S_4 = \mathcal{O}(N_x^{\alpha} L^{\beta})$, using the Upwind, Preissman Box and Lax-Wendroff schemes, given to 4dp, for $h = 0.5$ and $\mu = 1$. The results for N_x were identified using fixed $L = 4$ ($\Delta t = \frac{1}{2N_x}$), considering N_x in the form $N_x = 3^{\gamma}$, where $\gamma = 2, \ldots, 7$. The results for L were identified using fixed $N_x = 3^7$ ($\Delta t = \frac{1}{2 \cdot 3^7}$), considering L in the form $L = 2^{\delta}$, where $\delta = 0, \ldots, 9$. The results displayed here are the orders of convergence for the largest values of N_x and L considered, respectively. The full table of orders of convergence can be found in Appendix B.4.*

Now consider S_4 for the MNIMC scheme. Table 4.5 provides the numerical orders of convergence for S_4 to zero, with respect to N_x and L.

r	MNIMC	
	α	β
0	-7.2761×10^{-15}	5.7945×10^{-1}
1	-2.0000	5.6191×10^{-3}
2	-4.0000	5.6191×10^{-3}
3	-6.0000	5.6191×10^{-3}
4	-8.0000	5.6191×10^{-3}
5	-10.0000	5.6191×10^{-3}
6	-12.0000	5.6191×10^{-3}
7	-14.0000	5.6191×10^{-3}

Table 4.5: *The numerical orders of convergence to zero, with respect to N_x and L, for $S_4 = \mathcal{O}(N_x^{\alpha} L^{\beta})$, using the MNIMC scheme, given to 4dp, for $h = 0.5$ and $\mu = 1$. The results for N_x were identified using fixed $L = 4$ ($\Delta t = \frac{1}{2N_x}$), considering N_x in the form $N_x = 3^{\gamma}$, where $\gamma = 2, \ldots, 7$. The results for L were identified using fixed $N_x = 3^7$ ($\Delta t = \frac{1}{2 \cdot 3^7}$), considering L in the form $L = 2^{\delta}$, where $\delta = 0, \ldots, 9$. The results displayed here are the orders of convergence for the largest values of N_x and L considered, respectively. The full table of orders of convergence can be found in Appendix B.4.*

The numerical results in Table 4.5 give that the numerical orders of convergence to zero for S_4 using the MNIMC scheme, have the same properties as discussed for S_4 when using the Upwind, Preissman Box and Lax-Wendroff schemes. In the case of the MNIMC scheme, we represent ξ_p using (4.40) for all $p = 1, \ldots, \frac{N_x+1}{2}$, so is independent of p and N_x. As a result, S_4 has the following analytical order of convergence for the MNIMC scheme to zero, as derived in Appendix B.4,

$$S_4 = \mathcal{O}(N_x^{-2r}). \tag{4.48}$$

This analysis matches the numerical order of convergence to zero with respect to N_x for S_4, given in Table 4.5. However, (4.48) does not capture the numerical behaviour of S_4 with respect to L for the MNIMC scheme.

The order of convergence of S_5

$$S_5 = \frac{1}{N_x^{2r+1}} \sum_{p=2}^{\frac{N_x+1}{2}} |1 - \nu_p| \xi_p$$

This summation is composed from the amplification factors and the bounds on the error in the 1D DFT of $u_0(x)$ and the aliasing error due to the MNIMC scheme, giving it an explicit dependence on the regularity (r) of $u_0(x)$ over $(0, 1)$ and the number of discretisation points (N_x) when considering full sets of observations. The numerical orders of convergence to zero for this summation, with respect to both N_x and L, are given in Table 4.6.

The numerical orders of convergence with respect to N_x in Table 4.6 are identical to those for S_3 and S_4 for each regularity and scheme, making $S_5 = \mathcal{O}(N_x^{-2r})$. This gives

that S_5 does not decay to zero for $r = 0$, but does decay to zero for higher regularity initial conditions, as N_x is increased. We also see that the order of convergence to zero for each scheme, with respect to L is independent of regularity and are small in magnitude. The numerical orders of convergence to zero with respect to L displayed in Table 4.6 have yet to finish converging as L is increased.

Representing $|1 - \nu_p|$ and ξ_p using (4.38) and (4.40) respectively, Appendix B.5.3 derives an analytical order of convergence to zero of S_5, for the Upwind scheme,

$$S_5 = \mathcal{O}(LN_x^{-2r}). \tag{4.49}$$

This analytical result matches the numerical orders of convergence with respect to N_x in Table 4.6 for the Upwind scheme. The analytical order of convergence with respect to L does not capture the numerical behaviour of S_5 with respect to L.

r	α			β		
	Upwind	Preissman Box	Lax-Wendroff	Upwind	Preissman Box	Lax-Wendroff
0	1.8190×10^{-15}	-3.1679×10^{-6}	-3.5047×10^{-7}	6.4024×10^{-2}	-3.0357×10^{-1}	-3.8556×10^{-3}
1	-2.0000	-2.0000	-2.0000	6.4024×10^{-2}	-3.0357×10^{-1}	-3.8556×10^{-3}
2	-4.0000	-4.0000	-4.0000	6.4024×10^{-2}	-3.0357×10^{-1}	-3.8556×10^{-3}
3	-6.0000	-6.0000	-6.0000	6.4024×10^{-2}	-3.0357×10^{-1}	-3.8556×10^{-3}
4	-8.0000	-8.0000	-8.0000	6.4024×10^{-2}	-3.0357×10^{-1}	-3.8556×10^{-3}
5	-10.0000	-10.0000	-10.0000	6.4024×10^{-2}	-3.0357×10^{-1}	-3.8556×10^{-3}
6	-12.0000	-12.0000	-12.0000	6.4024×10^{-2}	-3.0357×10^{-1}	-3.8556×10^{-3}
7	-14.0000	-14.0000	-14.0000	6.4024×10^{-2}	-3.0357×10^{-1}	-3.8556×10^{-3}

Table 4.6: *The numerical orders of convergence to zero, with respect to N_x and L, for $S_5 = \mathcal{O}(N_x^\alpha L^\beta)$, using the Upwind, Preissman Box and Lax-Wendroff schemes, given to 4dp, for $h = 0.5$ and $\mu = 1$. The results for N_x were identified using fixed $L = 4$ ($\Delta t = \frac{1}{2N_x}$), considering N_x in the form $N_x = 3^\gamma$, where $\gamma = 2, \ldots, 7$. The results for L were identified using fixed $N_x = 3^7$ ($\Delta t = \frac{1}{2 \cdot 3^7}$), considering L in the form $L = 2^\delta$, where $\delta = 0, \ldots, 9$. The results displayed here are the orders of convergence for the largest values of N_x and L considered, respectively. The full table of orders of convergence can be found in Appendix B.5.*

The order of convergence of S_6

$$S_6 = \frac{1}{N_x^r} \sum_{p=2}^{\frac{N_x+1}{2}} \frac{|1-\nu_p|\xi_p}{(p-1)^{r+1}}$$

This summation is composed from the amplification factors and the bound on the error in the 1D DFT of $u_0(x)$, giving it an explicit dependence on the regularity (r) of $u_0(x)$ over $(0,1)$. The numerical orders of convergence to zero for this summation, with respect to both N_x and L, are given in Table 4.7.

Examining the numerical orders of convergence with respect to N_x for S_6 in Table 4.7, we see that S_6 is $\mathcal{O}(N_x^{-2r})$ for the Upwind scheme when $r = 0, 1$ and for the Preissman Box and Lax-Wendroff schemes when $r = 0, 1, 2$. Then the order of convergence changes to become $\mathcal{O}(N_x^{-r-2})$ for the Upwind scheme when $r \geq 3$ and $\mathcal{O}(N_x^{-r-3})$ for the Preissman Box and Lax-Wendroff schemes when $r \geq 4$. When $r = 2$ for the Upwind scheme and $r = 3$ for the Preissman Box and Lax-Wendroff schemes, the order of convergence to zero with respect to N_x is not obvious from Table 4.7. This can be explained for the Upwind scheme by considering the analytical order of convergence for S_6 to zero derived in Appendix B.6.3, by representing $|1-\nu_p|$ and ξ_p using (4.38) and (4.40) respectively.

$$S_6 = \begin{cases} \mathcal{O}(LN_x^{-2r}), & \text{for } r = 0, 1 \\ \mathcal{O}(LN_x^{-2r}\log(N_x)), & \text{for } r = 2, \\ \mathcal{O}(LN_x^{-r-2}), & \text{for } r \geq 3. \end{cases} \tag{4.50}$$

These analytical orders of convergence to zero match the numerical results for the numerical orders of convergence to zero with respect to N_x in Table 4.7 for the Upwind scheme. They show that similarly to S_2, when $r = 2$ a $\log(N_x)$ factor affects the order of convergence with respect to N_x. A similar factor may also be affecting the Preissman Box and Lax-Wendroff schemes at $r = 3$.

The numerical orders of convergence to zero with respect to L for S_6, vary with respect to regularity. They appear to be saturating to $\mathcal{O}(L)$ for a sufficiently large regularity. The saturated order of convergence to zero matches our analytical order of convergence to zero for the Upwind scheme with respect to L. However, this does not capture the numerical behaviour of S_6 with respect to L for small regularities.

r	α			β		
	Upwind	Preissman Box	Lax-Wendroff	Upwind	Preissman Box	Lax-Wendroff
0	2.0581×10^{-6}	-2.0772×10^{-6}	-5.8297×10^{-8}	2.3124×10^{-1}	1.4064×10^{-2}	7.6528×10^{-2}
1	-1.9988	-2.0000	-2.0000	5.6353×10^{-1}	3.4005×10^{-1}	3.0256×10^{-1}
2	-3.8520	-3.9991	-4.0002	9.2757×10^{-1}	6.6883×10^{-1}	6.3466×10^{-1}
3	-4.9993	-5.8571	-5.9305	1.0044	9.4645×10^{-1}	9.3594×10^{-1}
4	-6.0001	-6.9992	-7.0009	1.0034	1.0018	1.0001
5	-7.0000	-8.0000	-8.0000	1.0031	1.0028	1.0026
6	-8.0000	-9.0000	-9.0000	1.0030	1.0028	1.0028
7	-9.0000	-10.0000	-10.0000	1.0030	1.0028	1.0028

Table 4.7: *The numerical orders of convergence to zero, with respect to N_x and L, for $S_6 = \mathcal{O}(N_x^\alpha L^\beta)$, using the Upwind, Preissman Box and Lax-Wendroff schemes, given to 4dp, for $h = 0.5$ and $\mu = 1$. The results for N_x were identified using fixed $L = 4$ ($\Delta t = \frac{1}{2N_x}$), considering N_x in the form $N_x = 3^\gamma$, where $\gamma = 2, \ldots, 7$ for the Upwind and Preissman Box schemes and $\gamma = 2, \ldots, 12$ for the Lax-Wendroff scheme. The results for L were identified using fixed $N_x = 3^7$ ($\Delta t = \frac{1}{2 \cdot 3^7}$), considering L in the form $L = 2^\delta$, where $\delta = 0, \ldots, 9$. The results displayed here are the orders of convergence for the largest values of N_x and L considered, respectively. The full table of orders of convergence can be found in Appendix B.6.*

4.3.4 Discussion of the numerical results for summations S_1 to S_6

Our numerical results for the orders of convergence to zero of S_1 to S_6, have shown clearly how these summations depend on N_x, for each regularity initial condition. However, their dependence on L, is not as easy to decipher. The analytical orders of convergence have matched our numerical orders of convergence with respect to N_x for the Upwind scheme, for each summation. However, this is not true with respect to L. The asymptotic expansions of $|1-\nu_p|$ and ξ_p in (4.38) and (4.40) respectively, do not demonstrate the behaviour of these variables with respect to L adequately, to allow us to capture the behaviour analytically. The form of these variables with respect to L, did not allow for a Taylor expansion with respect to L about $L=0$, to be constructed. Another method is required to provide an asymptotic expansion of $|1-\nu_p|$ and ξ_p, that demonstrates their dependence on L, accurately.

In the case of the Preissman Box and Lax-Wendroff schemes, we did not have a bound for $|1-\nu_p|$ and ξ_p that allowed us to derive analytical bounds for the orders of convergence of summations S_1 to S_6. However, the numerical results for the three schemes show that there are similarities in the pattern of behaviour for all three schemes. Therefore it may be possible to use the form of $|1-\nu_p|$ and ξ_p in (4.38) and (4.40) respectively, together with how they relate to their numerical orders of convergence as demonstrated using the Upwind scheme, to suggest possible representations for these variables when considering the Preissman Box and Lax-Wendroff schemes. We now use these forms and the knowledge that $|1-\nu_p|$ appears to be $\mathcal{O}(LN_x^{-3})$ for the Preissman Box and Lax-Wendroff schemes when p is small in comparison to N_x ($p \neq 1$), from the numerical experiments in Section 4.3.1 and trial,

$$\begin{aligned} |1-\nu_p| &= \mathcal{O}\left(L\left[\tfrac{p-1}{N_x}\right]^3\right), \\ \xi_p &= \mathcal{O}(1). \end{aligned} \tag{4.51}$$

Substituting (4.51) into S_1 to S_6, we find that we achieve analytical orders of convergence to zero with respect to N_x, which match those in Tables 4.1-4.7. We also possess the same limitation as we did for the Upwind scheme. We are only able to match the orders of convergence to zero with respect to L, once they have reached saturation point, showing a limitation in our understanding of $|1-\nu_p|$ and ξ_p with respect to L.

When generating the numerical orders of convergence using the Lax-Wendroff scheme for S_1, S_2 and S_6 with respect to N_x, a larger range of values for N_x was required. This was because it took longer for the numerical results to finish converging. In fact the numerical results began converging to an alternative order of convergence to that displayed in Tables 4.1, 4.2 and 4.7, before reaching a critical value of N_x, where the order of convergence converged to those listed in the Tables. This can be seen in the results of Appendix B. This change was not displayed by any of the other schemes. How-

ever, since the numerically dissipative and non-dispersive Upwind and the numerically non-dissipative and dispersive Preissman Box schemes with respect to the resolvable wavenumber components of the numerical solution, have quite different orders of convergence, it could be that the combination of numerical dissipation and dispersion in the Lax-Wendroff scheme with respect to the resolvable wavenumber components, is causing a switch between the dominance of each form of error. It may also be linked to the $(p-1)^{-(r+1)}$ term found in S_1, S_2 and S_6, but not in any other summations.

4.3.5 The dominant summation

Now we have examined the order of convergence to zero with respect to both N_x and L for the six summations that make up the bound in Equation (4.19), we are able to determine which is the dominant summation with respect to N_x and L, for each regularity initial condition and scheme. The dominant summation in each instance can then be used to determine the order of convergence of the bound.

The summation with the dominant order of convergence to zero for the Upwind, Preissman Box and Lax-Wendroff schemes and any regularity initial condition, with respect to both N_x and L, was found to be S_1. This indicates that the errors introduced through numerical dissipation or dispersion of the resolvable wavenumber components of the numerical solution, have a greater influence on the error than aliasing errors. In the case of a numerically non-dissipative and non-dispersive scheme with respect to the resolvable wavenumber components of the solution, S_4 is the only summation in the bound, so this is the dominant summation in this instance. This is the case for the MNIMC scheme.

The next step in the analysis is to compare the numerical order of convergence of the bound in Equation (4.19) to zero, with the order of convergence of the strong constraint 4D-Var numerical experiments to zero, performed in Section 4.3. The order of convergence for the bound in Equation (4.19) is given by the order of convergence of the dominant summation for the relevant scheme. As a result, we compare the orders of convergence for the relevant dominant summation to zero, with those from the strong constraint 4D-Var numerical experiments to zero.

These numerical experiments included a Gaussian initial condition. As discussed in Section 4.2.1, large r will be considered in this instance. The analysis of S_1 in Section 4.3.1 suggests that the order of convergence of S_1 saturates at $\mathcal{O}(N_x^{-3})$ when $r \geq 2$, for the Upwind scheme and $\mathcal{O}(N_x^{-5})$ when $r \geq 3$, for the Preissman Box and Lax-Wendroff schemes. In order to verify this, an asymptotic analysis of S_1 is performed to determine its leading order behaviour with respect to N_x as $N_x \to \infty$.

Using Equation (4.38) we find that for large r,

$$\sum_{p=2}^{\frac{N_x+1}{2}} \frac{|1-\nu_p|^2}{(p-1)^{2(r+1)}} \sim |1-\nu_2|^2, \text{ as } N_x \to \infty.$$
$$\Rightarrow \quad S_1 \sim N_x|1-\nu_2|^2, \text{ as } N_x \to \infty. \tag{4.52}$$

This leading order behaviour is due to the rapid decay of $(p-1)^{-2(r+1)}$ to zero for large r, when $p = 3, \ldots, \frac{N_x+1}{2}$, as $N_x \to \infty$. When $p = 2$, $(p-1)^{-2(r+1)}$ remains constant for any value of r, hence $|1-\nu_2|^2$ determines the behaviour of S_1 for large r. When considering the order of convergence of S_1 for large r, the order of convergence to zero of (4.52) will be considered. This is given by,

$$S_1 \sim N_x|1-\nu_2|^2 = \mathcal{O}(N_x^{1+2\gamma}), \tag{4.53}$$

where γ is the order of convergence of $|1-\nu_p|$ to zero with respect to N_x, for small p, $p \neq 1$, determined numerically in Section 4.3.1 eg. for the Upwind scheme $\gamma = -2$. This gives S_1 an order of convergence $\mathcal{O}(N_x^{-3})$ for the Upwind scheme and $\mathcal{O}(N_x^{-5})$ for the Preissman Box and Lax-Wendroff schemes. These orders of convergence are equal to the saturated orders of convergence found for S_1 numerically, for each scheme in Section 4.3.1. Since the saturated orders of convergence were found for relatively small values of r, it shows that the leading order behaviour of S_1 is dominant for relatively small values of r.

Figures 4.1 and 4.2 also contain the plots of $2D_2^2S_1$ for each initial condition considered in the strong constraint 4D-Var data assimilation numerical experiments of the Section. The figure shows that for each initial condition and finite difference scheme considered, $2D_2^2S_1$ formed a tighter bound than (4.19), for the error in the analysis vector. It also appears to possess the same orders of convergence to zero as N_x and L are increased. $2D_2^2S_1$ most likely forms a tighter bound due to the use of the triangle inequality in the proof of Lemma 4.4. The error due to aliasing initially began as $-\boldsymbol{\rho}_L$ in the proof. By using the triangle inequality, the summations S_4, S_5 and S_6 relating to $\boldsymbol{\rho}_L$ became positive quantities added rather than subtracted from the bound, increasing its value.

In the case of a numerically non-dissipative and non-dispersive scheme with respect to the resolvable wavenumber components of the numerical solution, ξ_p is constant with respect to p and N_x. As a result $\xi_p = \xi_1$ for all $p = 1, \ldots, N_x$, hence,

$$S_4 = \frac{\xi_1^2}{N_x^{2r}}. \tag{4.54}$$

4.3.6 Comparison of numerical orders of convergence

We now compare the numerical orders of convergence for the relevant dominant summations identified in Section 4.3.5, with those found during the strong constraint 4D-Var data assimilation numerical experiments of Section 4.3. Table 4.8 displays the orders of convergence to zero for the l_2-norm of the error in the analysis vector found through strong constraint 4D-Var data assimilation numerical experiments, for differing regularity initial conditions. The results presented in the table are for the largest considered value of N_x and L. Figures 4.4 and 4.5 plot the numerical results for all values of N_x and L considered, respectively. These results should be compared with the numerical orders of convergence for S_1 for the Upwind, Preissman Box and Lax-Wendroff schemes, found in Table 4.1. The numerical orders of convergence for the MNIMC scheme in Table 4.8, should be compared with those in Table 4.5, for summation S_4.

Variable	Upwind Scheme		
	1D Square Function ($r = 0$)	Triangular Function ($r = 1$)	1D Gaussian Function ($r \gg 1$)
α	1.1838×10^{-12}	-2.2612	-3.0000
β	5.6939×10^{-1}	1.5096	2.0000
Variable	Preissman Box Scheme		
	1D Square Function ($r = 0$)	Triangular Function ($r = 1$)	1D Gaussian Function ($r \gg 1$)
α	-6.5427×10^{-1}	-1.2809	-4.9178
β	3.7952×10^{-1}	9.8836×10^{-1}	2.0662
Variable	Lax-Wendroff Scheme		
	1D Square Function ($r = 0$)	Triangular Function ($r = 1$)	1D Gaussian Function ($r \gg 1$)
α	5.5724×10^{-1}	-2.0836	-4.9947
β	3.1248×10^{-1}	1.0187	2.0194
Variable	MNIMC Scheme		
	1D Square Function ($r = 0$)	Triangular Function ($r = 1$)	1D Gaussian Function ($r \gg 1$)
α	-2.7652×10^{-3}	-1.9984	-2.0602
β	5.6191×10^{-3}	5.6191×10^{-3}	5.2940

Table 4.8: *Numerical orders of convergence to zero for $\|\tilde{\mathbf{x}}_0 - \mathbf{x}_a\|_2^2$, with respect to N_x and L, for the l_2-norm of the error in the analysis vector from strong constraint 4D-Var experiments, given to 4dp, $\|\tilde{\mathbf{x}}_0 - \mathbf{x}_a\|_2^2 = \mathcal{O}(N_x^\alpha L^\beta)$, with $h = 0.5$ and $\mu = 1$. The results for N_x and L were identified using fixed $L = 4$ ($\Delta t = \frac{1}{2N_x}$) and fixed $N_x = 3^7$ ($\Delta t = \frac{1}{2 \cdot 3^7}$), respectively. The results displayed here are the orders of convergence for the largest values of N_x and L considered, respectively.*

Initially consider the results for the order of convergence to zero with respect to N_x in Table 4.8 for the Upwind, Preissman Box and Lax-Wendroff schemes. The results are close to those in Table 4.1 for initial conditions with the same value of r. Figure 4.4 plots the numerical order of convergence with respect to N_x, as N_x is increased in powers of three. It shows that the order of convergence fluctuates about the order of convergence shown in Table 4.1, for each value of r. This explains why the results in Tables 4.1 and 4.8 do not match exactly for N_x. The trends of the plots in Figure 4.4

show that the numerical orders of convergence for S_1 in Table 4.1 are a good match to those from the strong constraint 4D-Var numerical experiments, with respect to N_x. Table 4.1 also shows that the order of convergence with respect to L is a good match to those found in Table 4.8. These results indicate that (4.19) is an appropriate bound for characterising the behaviour of the l_2-norm of the error in the analysis vector, with respect to N_x and L, for the Upwind, Preissman Box and Lax-Wendroff schemes. This behaviour can be identified by examining $2D_2^2 S_1$.

Next we consider the order of convergence with respect to N_x in Table 4.8, for the MNIMC scheme. The results for $r = 0$ and $r = 1$ in Table 4.5 are close to those of Table 4.8 for the square and triangular functions respectively. However, we notice that in Table 4.8, that the 1D Gaussian function has an order of convergence of $\mathcal{O}(N_x^{-2})$. This appears to show that the order of convergence with respect to N_x of the l_2-norm of the error in the analysis vector for the MNIMC scheme, saturates once a critical regularity has been reached, as the 1D Gaussian function corresponds to large r. Table 4.5 does not shows that S_4 to have this property. Instead S_4 has an order of convergence to zero of $\mathcal{O}(N_x^{-2r})$. Therefore analysing S_4 is not sufficient to characterise the behaviour of the l_2-norm of the error in the analysis vector, nor does it appear to form a bound for this error when $r > 1$.

Comparing the numerical orders of convergence to zero for S_4 in Table 4.5 with those in Table 4.8 for the strong constraint 4D-Var numerical experiments, with respect to L for the MNIMC scheme, we find the same result. S_4 characterises the behaviour of the l_2-norm of the error in the analysis vector for $r = 0, 1$, but not for higher regularities.

Comparing the orders of convergence for each scheme in Figures 4.4 and 4.5, we notice that the order of convergence with respect to N_x for the 1D square function is $\mathcal{O}(N_x^0)$. The relevant dominant summation for each scheme, indicate that this would be the order of convergence to zero for these schemes, given any initial condition with regularity zero. This indicates that the error in the analysis vector does not decay as N_x is increased. This is most likely due to the error that always exists when a Fourier series is used to approximate a discontinuous function, as the Fourier series converges to the midpoint of the jump discontinuity.

Higher regularity initial conditions possess an order of convergence to zero, with respect to N_x, which is less than zero. This indicates that the error in the analysis vector decreases as the number of discretisation points when considering full sets of observations, is increased. As discussed in Section 4.3, this is equivalent to the error in the analysis vector decreasing as the density of observations in space and time is increased. This is consistent with our intuition that increasing the number of discretisation points will decrease the error in the analysis vector and with the results of Rabier et al. [4], who found they gained improvements in the results from 3D-Var and incremental 4D-Var experiments conducted at a higher grid resolution.

The error in the results for the MNIMC scheme for the triangular function and 1D Gaussian function initial conditions, is smaller than for the other schemes for the

considered values of N_x and L. However the order of convergence with respect to N_x is faster for the 1D Gaussian function initial condition, for the numerically dissipative and/or dispersive schemes with respect to the resolvable wavenumber components of the numerical solution. Therefore for sufficiently large N_x, the error introduced by the numerically dissipative and/or dispersive schemes with respect to the resolvable wavenumber components, will be smaller than the error introduced by the MNIMC scheme. This may indicate that when considering a numerical model with a large number of discretisation points when considering full sets of observations, a numerically dissipative and/or dispersive scheme with respect to the resolvable wavenumber components of the numerical solution may reduce the effects of numerical model error.

Examining Figure 4.5, we notice that the numerically dissipative and/or dispersive schemes with respect to the resolvable wavenumber components, all have orders of convergence to zero with respect to L, that show that the error in the analysis vector increases as the number of sets of observations in the assimilation window is increased. This is an unexpected result. We also get this result when considering the MNIMC scheme for the 1D Gaussian function initial condition with respect to L. However, the error introduced by the scheme remains constant with respect to L for the square and triangular function initial conditions. This needs further research to understand this behaviour.

Consider the orders of convergence to zero for the l_2-norm of the error in the analysis vector in (4.13), found through the truncation error method in Lemma 4.1. We can see that the order of convergence to zero for the Upwind, Preissman Box and Lax-Wendroff schemes with respect to N_x, are the saturated orders of convergence identified for S_1 and the strong constraint 4D-Var numerical experiments in Figure 4.5 with respect to N_x. The saturated orders of convergence are for higher regularity initial conditions. This agrees with the conditions of Lemma 4.1 that (4.13) only holds for sufficiently smooth initial conditions. The orders of convergence to zero for S_1 and the results from the strong constraint 4D-Var numerical experiments, with respect to L are not captured by (4.13). Comparing the bound in (4.13) with (4.19) with respect to L, we see that (4.19) forms a better approximation for the behaviour of the error with respect to L, even though its not perfect.

In this Section, we have examined the numerical orders of convergence for initial conditions, whose large scale behaviour demonstrates its regularity. We now examine an initial condition whose large scale behaviour does not match its regularity and investigate how the behaviour of the bound in (4.19) and the error found through numerical experiments, differ. This analysis allows us to interpret the information being provided by the bound, more clearly.

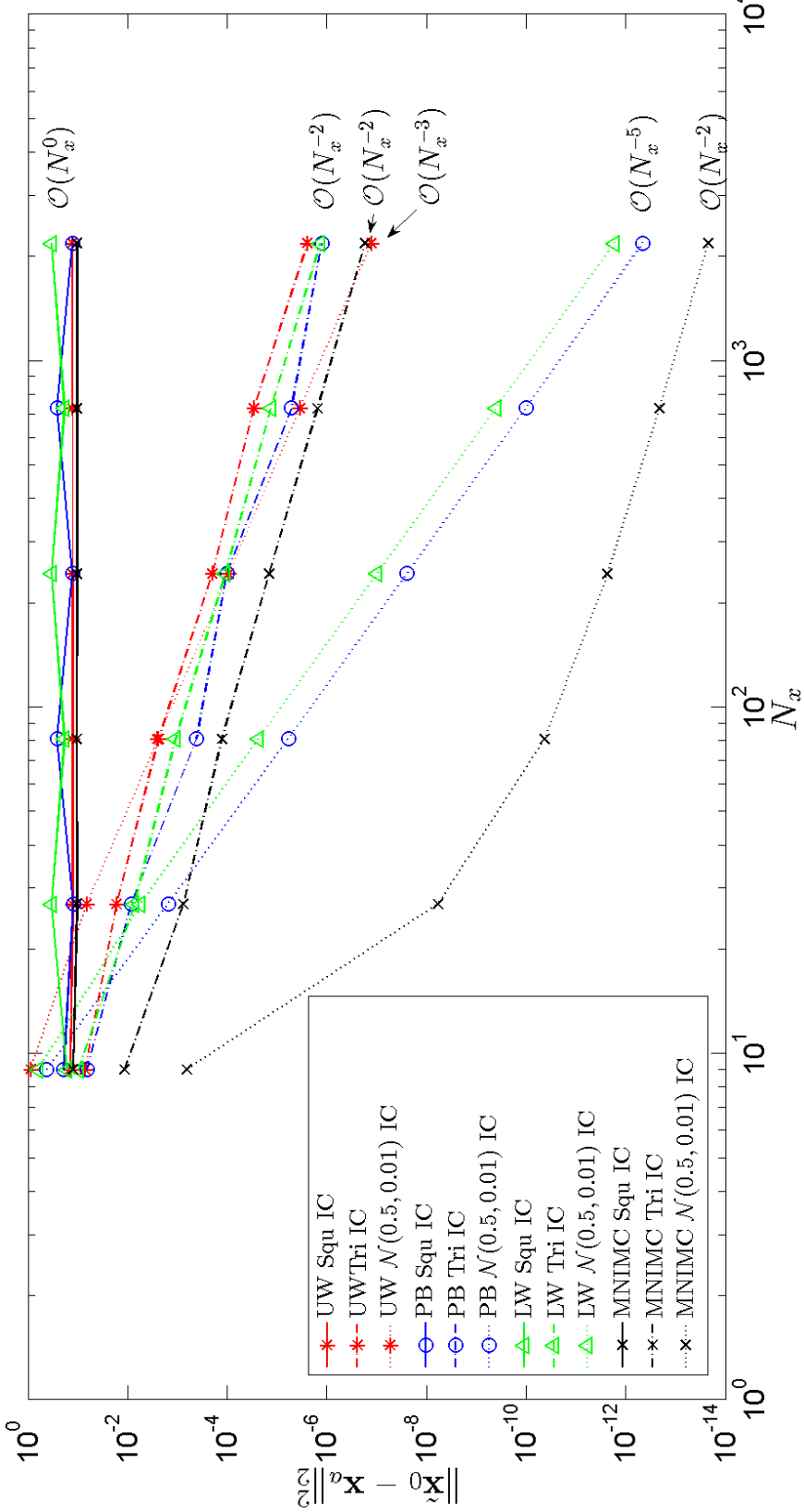


Figure 4.4: *The square of the l_2-norm of the error in the analysis vector, calculated through strong constraint 4D-Var data assimilation numerical experiments, solely under the influence of errors introduced by finite difference approximations in the forward model. The results were generated using the Upwind (UW), Preissman Box (PB), Lax-Wendroff (LW) and MNIMC schemes as the forward models for solving the 1D linear advection problem in (3.1), using $h = 0.5$, $\mu = 1$, $L = 4$ and $N_x = 3^\gamma$ where $\gamma = 2, \ldots, 7$, ($\Delta t = \frac{1}{2N_x}$). The functions considered for $u_0(x)$ in these experiments are defined in Section 4.3, where 'squ IC', 'tri IC', $\mathcal{N}(0.5, 0.01)$ IC' denote the 1D square, the triangular and 1D Gaussian functions respectively. The results are plotted using logarithmic scales to demonstrate the order of convergence of the error to zero, with respect to N_x.*

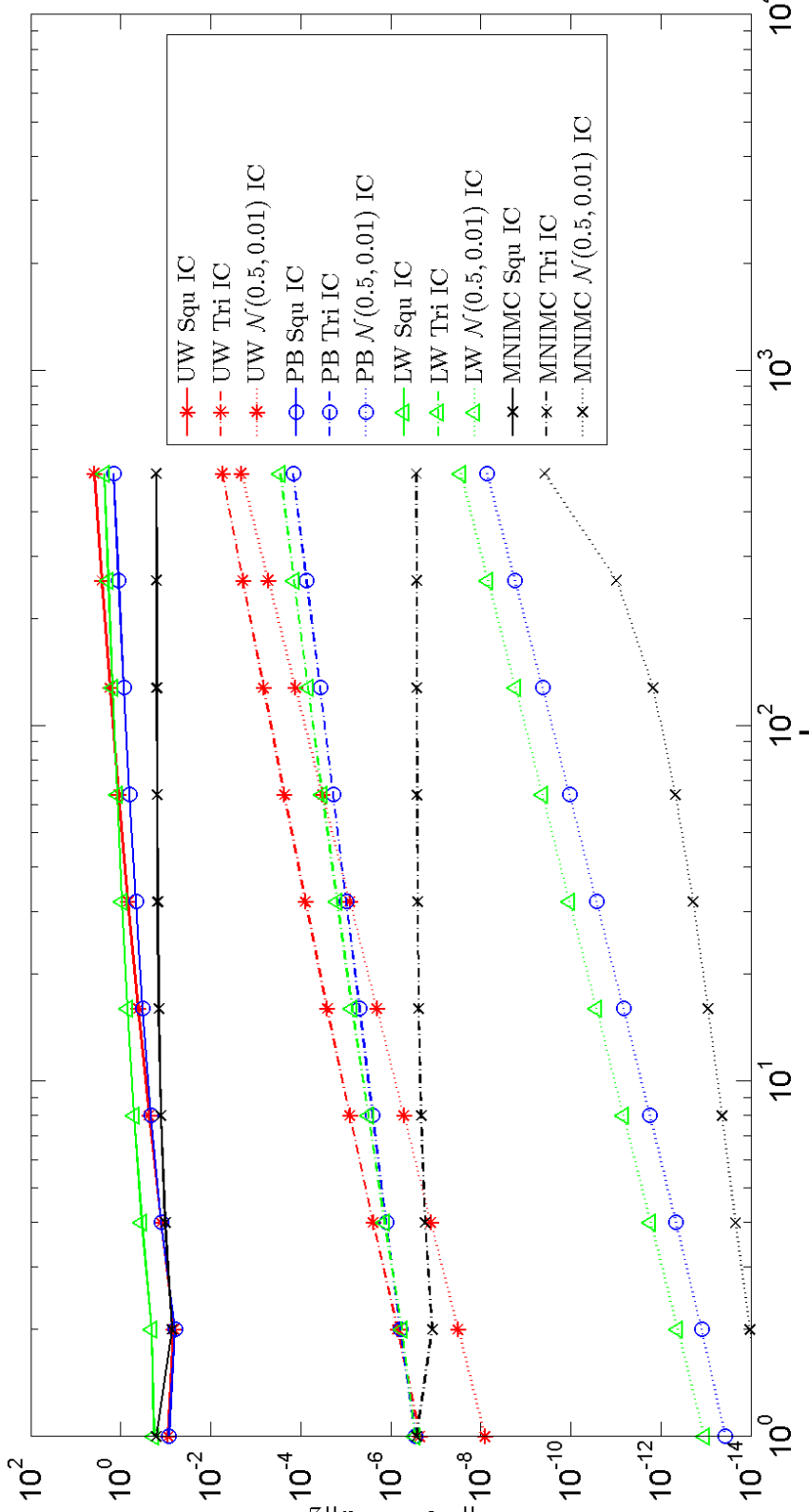


Figure 4.5: *The square of the l_2-norm of the error in the analysis vector, calculated through strong constraint 4D-Var data assimilation numerical experiments, solely under the influence of errors introduced by finite difference approximations in the forward model. The results were generated using the Upwind (UW), Preissman Box (PB), Lax-Wendroff (LW) and MNIMC schemes as the forward models for solving the 1D linear advection problem in (3.1), using $h = 0.5$, $\mu = 1$, $N_x = 3^7$ and $L = 2^\delta$ where $\delta = 0, \ldots, 9$, ($\Delta t = \frac{1}{2 \cdot 3^7}$). The functions considered for $u_0(x)$ in these experiments are defined in Section 4.3, where 'squ IC', 'tri IC', $\mathcal{N}(0.5, 0.01)$ IC' denote the 1D square, the triangular and 1D Gaussian functions respectively. The results are plotted using logarithmic scales to demonstrate the order of convergence of the error to zero, with respect to L.*

4.3.7 Interpreting the bound on the error in the analysis vector

Examining the proof of Lemma 4.4, we see that the corresponding Fourier coefficient of $u_0(x)$ is added and subtracted from $\mathcal{F}_p(\tilde{\mathbf{x}}_0)$, in (4.23). Notice though that the Fourier coefficient of any function that is both continuous and has the same values as $u_0(x)$ at the sample points and possesses a convergent Fourier series, could have been used. Selecting such a function and increasing N_x means that as the two functions are different, there will be a critical value of N_x at which this function no longer satisfies the above criteria at all discretisation points. At this point a new function satisfying the above criteria for this critical value of N_x could then be used. The pool of functions available to do this will shrink as N_x increases. The only function that will consistently provide Fourier coefficients which can be used as in (4.23), is the function $u_0(x)$ from which the discrete sample was taken. Hence the proof makes use of the Fourier coefficients for this function.

Suppose for a moment that we select a function satisfying the above requirements and use its Fourier coefficients in (4.23), rather than those of $u_0(x)$. Then the order of convergence of this bound will be determined by the regularity of this function. The regularity of this function may be higher than the regularity of $u_0(x)$. In which case the order of convergence to zero for the bound in (4.19) may be higher than the order of convergence of the l_2-norm of the error in the analysis vector. As N_x is increased and we reach the critical value of N_x, the regularity of our new choice of function will determine the order of convergence of our bound in Equation (4.19). Eventually, the regularity of $u_0(x)$ will determine the order of convergence of the bound. As a result, the best option is to choose to use the function $u_0(x)$. However, it may be that the behaviour of the numerical results corresponds to the regularity of the initial condition that best represents the discrete observations. In order to explore this, consider the following function,

$$u_0(x) = \begin{cases} \frac{10}{\sqrt{2\pi}} e^{-50\left(x-\frac{1}{2}\right)^2}, & \text{for } x \in [0,1) \setminus \left(\frac{300}{3^6}, \frac{301}{3^6}\right), \\ 0.2, & \text{for } x \in \left(\frac{300}{3^6}, \frac{301}{3^6}\right), \end{cases} \tag{4.55}$$

This is the 1D Gaussian function in (4.31) with a step cut into it, as can be seen in Figure 4.6. This function has $v_1 = v_2 = \frac{10}{\sqrt{2\pi}}$, $s = 2$ and

$$w = \begin{cases} 1, & \text{for } \frac{N_x}{5} \notin \mathbb{N}, \\ 2, & \text{for } \frac{N_x}{5} \in \mathbb{N}. \end{cases}$$

As we are considering $N_x = 3^\alpha$ for $\alpha \in \mathbb{N}$ in our numerical strong constraint 4D-Var experiments, we have $w = 1$. This results in $D_1 = \frac{10}{\sqrt{2\pi}}$, $D_2 = \frac{40}{\sqrt{2\pi}^{\frac{3}{2}}}$ and $D_3 = 2D_2[2 + \zeta(2)] + 2v_1$.

When N_x is sufficiently small, the cut will be sub-grid scale, so observations of the function will make the function appear to be a Gaussian function. As N_x increases, a

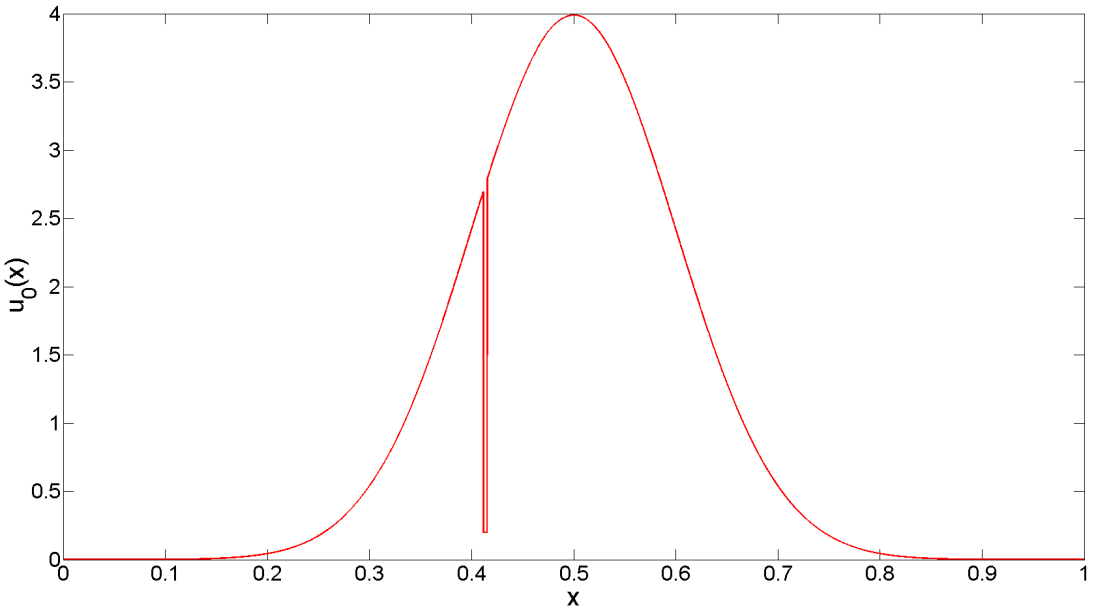


Figure 4.6: *The function $u_0(x)$ in (4.55).*

critical value of N_x will be reached where observations are taken of the step cut into the Gaussian function. As we are considering N_x in the form of powers of three in our numerical experiments, observations of the step will only occur when $\Delta t < \frac{1}{3^6}$, as observations only see the function sampled every $\frac{\Delta x}{2}$ in time (as $h = 0.5$). Then when $N_x = 3^6$, observations will be taken of the step for the first time.

Strong constraint 4D-Var data assimilation numerical experiments were conducted for the function in (4.55), under the conditions set out in Section 4.3. Figure 4.7 shows how the error in the analysis vector described by the bound in (4.19) behaves with respect to N_x, in comparison to the same for the strong constraint 4D-Var numerical experiments, for the Upwind scheme. Here we see that the bound and the dominant summation for the scheme have an order of convergence to zero of $\mathcal{O}(N_x^0)$, for an initial condition whose regularity is $r = 0$, as would be expected from our previous analysis for the 1D square function. However, the strong constraint 4D-Var numerical results initially show an order of convergence $\mathcal{O}(N_x^{-3})$, as would be expected for the 1D Gaussian function. Once a critical value of N_x has been reached, the order of convergence becomes $\mathcal{O}(N_x^0)$. This critical value is $N_x = 3^6$. This is the point we expected the switch to occur as the observations are now able to observe the small scale behaviour of the function. This in fact shows that we could have used the higher regularity functions in the bound up until this point. However, the wisest course of action is to use the order of convergence provided by $u_0(x)$, as eventually the error in the numerical experiments will match this. In this way, the bound provides the worst case behaviour of the l_2-norm of the error in the analysis vector.

Figures 4.8(a) and 4.8(b) show the order of convergence to zero of the error in the analysis vector, with respect to N_x and L respectively, as found through strong

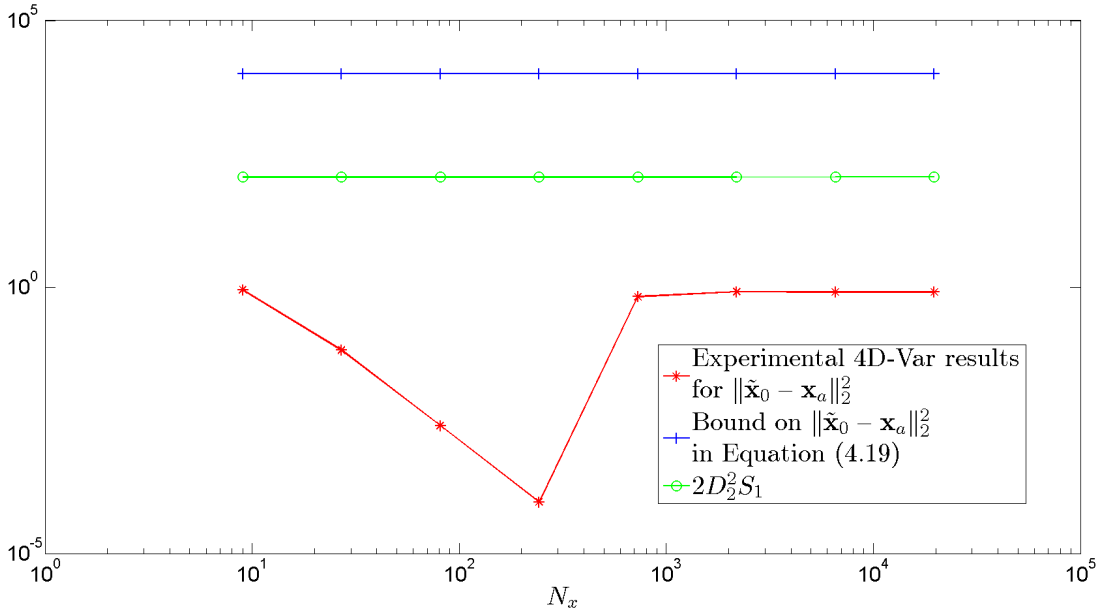
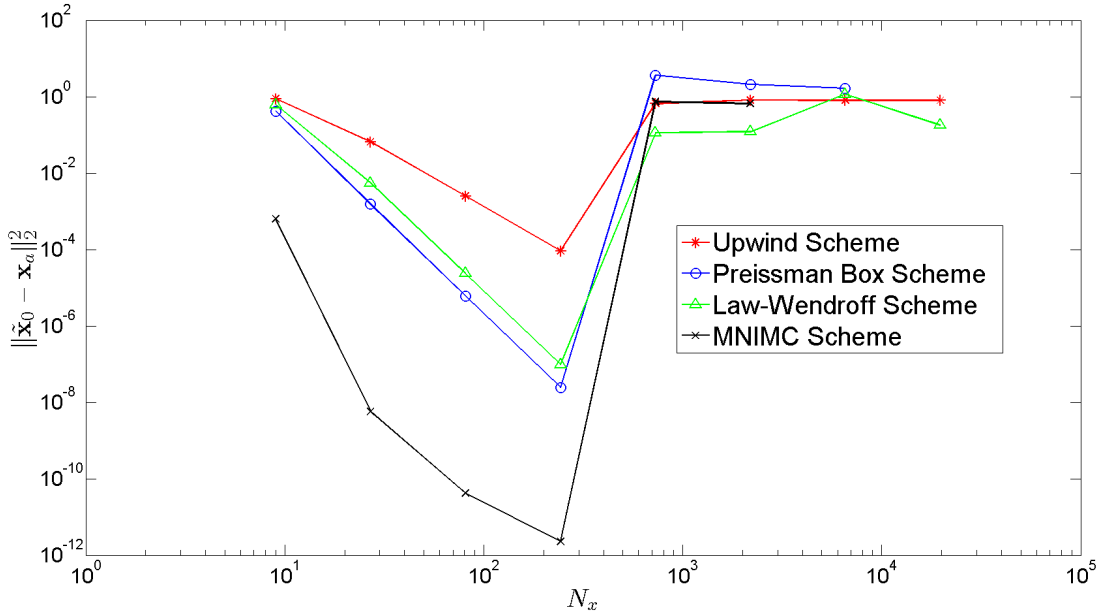


Figure 4.7: *The square of the l_2-norm of the error in the analysis vector, calculated through strong constraint 4D-Var numerical experimentation, solely under the influence of errors introduced by finite difference approximations in the forward model. The results were generated using the Upwind scheme for solving the 1D linear advection problem in (3.1), using $h = 0.5$, $\mu = 1$, $L = 4$ and $N_x = 3^\gamma$ for $\gamma = 3, \ldots, 9$. The considered $u_0(x)$ for the true initial condition is (4.55). The bound for the error in Equation (4.19) and its dominant summation $2D_2^2S_1$ for the considered scheme, are plotted alongside for comparison, using the same variables. The results are plotted using logarithmic scales to demonstrate the order of convergence with respect to N_x, of the error to zero.*

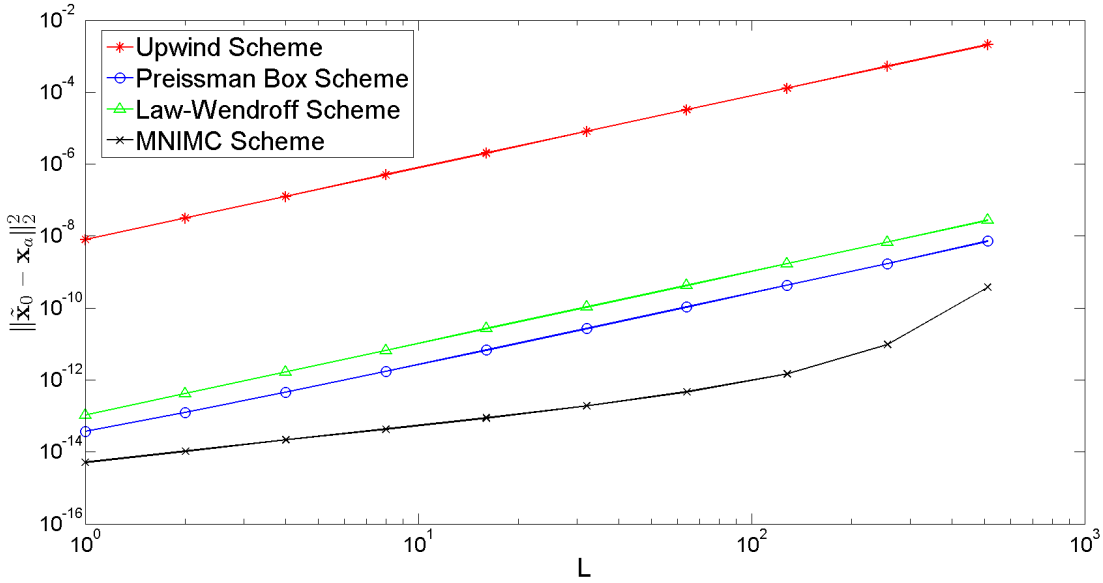
constraint 4D-Var numerical experiments. Considering Figure 4.8(a), we see that the order of convergence to zero with respect to N_x for each scheme, is initially identical to that found for the 1D Gaussian function in the previous Section. Once the critical value of N_x has been reached, the order of convergence for each scheme changes to that of an initial condition with regularity $r = 0$.

Examining Figure 4.8(b), we see that the orders of convergence with respect to L, do not experience a similar change. This is because observations are taken at every point in space and time, so as N_x is fixed, the observation points of the function are fixed. Increasing L only increases the length of the observation window. Even though $N_x = 3^7$, so is large enough to observe the small scale behaviour of (4.55), the numerical errors for each scheme behave as for the 1D Gaussian function in Section 4.3.5. This is unexpected and requires further research to see if increasing N_x does allow the order of convergence with respect to L to change to that of a function with regularity $r = 0$.

In reality it is not possible to remove all forms of error other than numerical model error, when implementing strong constraint 4D-Var data assimilation, as we have done so here. Other forms of error interact with the numerical model error, to affect its behaviour. Now we have some insight into the effects of numerical model error, we will now re-introduce other forms of error to the problem, to observe the effect of their



(a) The results for fixed $L = 4$, where $N_x = 3^\gamma$ for $\gamma = 2, \ldots, 9$.



(b) The results for fixed $N_x = 3^7$, where $L = 2^\delta$ for $\delta = 0, \ldots, 9$.

Figure 4.8: *The square of the l_2-norm of the error in the analysis vector, calculated through strong constraint 4D-Var numerical experimentation, solely under the influence of errors introduced by finite difference approximations in the forward model. The results were generated using the Upwind, Preissman Box, Lax-Wendroff and MNIMC schemes for solving the 1D linear advection problem in (3.1), using $h = 0.5$, $\mu = 1$, $L = 4$. The considered function for $u_0(x)$ in these experiments is (4.55). The results are plotted using logarithmic scales to demonstrate the order of convergence with respect to N_x in Figure 4.8(a) and L in Figure 4.8(b), of the error to zero.*

combinations. We choose to re-introduce observation errors as the forecast from the analysis vector has been shown to be most sensitive to changes in observation errors [3]. We derive a similar bound to that derived in Lemma 4.4 for this problem, as it was

a good fit for depicting the order of convergence of the error in the analysis vector with respect to N_x, for the Upwind, Preissman Box and Lax-Wendroff schemes. We then analyse the order of convergence of this bound to zero and compare it with the order of convergence to zero, for the results of strong constraint 4D-Var numerical experiments, with respect to both N_x and L. We consider this problem in the next Section.

4.4 Spectral approach with observation errors

Section 4.2 provided a bound for the l_2-norm of the error in the analysis vector, due to numerical model error introduced by finite difference approximations in the forward model, for different regularity initial conditions. It is possible to develop a similar bound for the l_2-norm of the error in the analysis vector when observation errors are also included.

Consider the case where each observation contains observation errors, $\mathbf{y}_l = \tilde{\mathbf{y}}_l + \boldsymbol{\epsilon}_l$, as described in Section 2.3. Specifically, let us consider the random error known as *white noise* [79] where $\boldsymbol{\epsilon}_l \sim \mathcal{N}(\mathbf{0}, \sigma_o^2 I_{N_x})$. The cost function is defined as in Equation (2.10),

$$J(\mathbf{x}_0) = \frac{1}{\sigma_o^2} \sum_{l=0}^{L} \left[\mathbf{y}_l - M^l \mathbf{x}_0\right]^T \left[\mathbf{y}_l - M^l \mathbf{x}_0\right]. \tag{4.56}$$

Minimising (4.56) with respect to $\mathbf{x}_0$, yields,

$$\mathbf{x}_a = \left[\sum_{k=0}^{L} (M^T M)^k\right]^{-1} \sum_{l=0}^{L} (M^T)^l \left[\tilde{\mathbf{y}}_l + \boldsymbol{\epsilon}_l\right]. \tag{4.57}$$

Using the eigenvalue decomposition of M and $\tilde{M}$ as well as Lemma 4.3 for finite L,

$$\mathbf{x}_a = A_L \tilde{\mathbf{x}}_0 + \boldsymbol{\rho}_L + V \left[\sum_{k=0}^{L} (\Lambda^* \Lambda)^k\right]^{-1} \left[\sum_{l=0}^{L} (\Lambda^*)^l V^* \boldsymbol{\epsilon}_l\right]. \tag{4.58}$$

The analysis vector in (4.58) is expressed in part by the analysis vector without observation errors, as in (3.70). The observation errors form a separate term. If the errors were not assumed to have the same variance, (4.58) would not have this property.

The term containing the observation errors in (4.58), would be the analysis vector when considering observations of the form $\mathbf{y}_l = \boldsymbol{\epsilon}_l$. The effect of numerical model error introduced by finite difference approximations, on the white noise observation errors, may lead to correlations within the observation noise component of (4.58). If correlations have been introduced, then this will create artifacts in the analysis vector which will be propagated into its forecast. The autocorrelation function is used to determine if the observation noise contribution to $\mathbf{x}_a$ is still white noise.

The autocorrelation function is defined as in Mitra [79] and Briggs [60]. The autocorrelation of an N_x-periodic sample $\mathbf{x} \in \mathbb{R}^{N_x}$, at lag $j = 0, \ldots, N_x - 1$, is defined as

$z_j : \mathbb{R}^{N_x} \to \mathbb{R}$, such that $\mathbf{x} \mapsto z_j(\mathbf{x})$ where,

$$z_j(\mathbf{x}) = \frac{1}{N_x} \sum_{p=1}^{N_x} \{\mathbf{x}\}_p \{\mathbf{x}\}_{[p-j]_{N_x}}, \tag{4.59}$$

where $\{\mathbf{x}\}_p$ denotes the pth element of $\mathbf{x}$ and $[\cdot]_{N_x}$ denotes modulo N_x. Also, define $\mathbf{z} \in \mathbb{R}^{N_x}$ such that the jth element of $\mathbf{z}$ is $z_{j-1}(\cdot)$. Then by the Wiener-Khintchine Theorem [79], the DFT of the autocorrelation of $\mathbf{x}$ is defined as,

$$\mathcal{F}[\mathbf{z}(\mathbf{x})] = \frac{1}{\sqrt{N_x}} \left[|\mathcal{F}_1(\mathbf{x})|^2, |\mathcal{F}_2(\mathbf{x})|^2, \ldots, |\mathcal{F}_{N_x}(\mathbf{x})|^2 \right]^T.$$

Using (4.58), the autocorrelation of the noise component of the analysis vector is given by,

$$z_{j-1}\left(V \left[\sum_{k=0}^{L} (\Lambda^* \Lambda)^k \right]^{-1} \sum_{l=0}^{L} (\Lambda^*)^l V^* \boldsymbol{\epsilon}_l \right) = \frac{1}{N_x} \sum_{p=1}^{N_x} \left| \frac{\sum_{l=0}^{L} \overline{\lambda}_p^l \mathcal{F}_p(\boldsymbol{\epsilon}_l)}{\sum_{k=0}^{L} |\lambda_p|^{2k}} \right|^2 e^{\frac{2\pi i (j-1)(p-1)}{N_x}}, \tag{4.60}$$

for $j = 1, \ldots, N_x$. Hence,

$$\mathbb{E}\left[z_{j-1}\left(V \left[\sum_{k=0}^{L} (\Lambda^* \Lambda)^k \right]^{-1} \sum_{l=0}^{L} (\Lambda^*)^l V^* \boldsymbol{\epsilon}_l \right) \right] = \frac{\sigma_o^2}{N_x} \sum_{p=1}^{N_x} \frac{e^{\frac{2\pi i (j-1)(p-1)}{N_x}}}{\sum_{k=0}^{L} |\lambda_p|^{2k}}, \quad \forall j = 1, \ldots, N_x, \tag{4.61}$$

which relies upon the values of j, N_x, L and σ_o^2, together with the dissipative properties of the considered finite difference scheme, with respect to the resolvable wavenumber components of the numerical solution. It does not utilise the dispersive properties of the scheme with respect to the resolvable wavenumber components of the numerical solution. Expression (4.61) is potentially non-zero for all j, for a numerically dissipative finite difference scheme, indicating that the noise component of the analysis vector may no longer be random white noise. However, in the case of a non-dissipative scheme ie: $|\lambda_p| = 1 \quad \forall p$, only $j = 1$ is non-zero. Using a non-dissipative scheme such as the Preissman Box scheme, means that the noise component of the analysis vector will retain the white noise structure implicit in the observations.

A spectral approach as in Section 4.2 is now used to provide a bound for the l_2-norm of the error in the analysis vector, for any regularity initial condition, in the presence of numerical model error due to finite difference approximations and observation errors.

Lemma 4.5. *Let the assumptions of Lemma 4.4 hold true but consider observations of the form* $\mathbf{y}_l := \tilde{\mathbf{y}}_l + \boldsymbol{\epsilon}_l$*, allowing* $\mathbf{x}_a$ *to be stated as in (3.70). Then,*

$$\|\tilde{\mathbf{x}}_0 - \mathbf{x}_a\|_2^2 \leq E_M + E_O + E_C, \tag{4.62}$$

where

$$E_M = N_x \left\{ |1-\nu_1| D_1 + (|1-\nu_1| + 2\xi_1) \frac{D_3}{N_x^{r+1}} \right\}^2 + 2N_x \sum_{p=2}^{\frac{N_x+1}{2}} \left\{ |1-\nu_p| \frac{D_2}{(p-1)^{r+1}} + (|1-\nu_p| + 2\xi_p) \frac{D_3}{N_x^{r+1}} \right\}^2, \tag{4.63}$$

$$E_O = N_x z_0 \left(\left[\sum_{k=0}^{L} (\Lambda^* \Lambda)^k \right]^{-1} \sum_{l=0}^{L} (\Lambda^*)^l V^* \boldsymbol{\epsilon}_l \right), \tag{4.64}$$

$$E_C = 2\sqrt{N_x} \left\{ |1-\nu_1| D_1 + (|1-\nu_1| + 2\xi_1) \frac{D_3}{N_x^{r+1}} \right\} \left| \frac{\sum_{l=0}^{L} \overline{\lambda}_1^l \mathcal{F}_1(\boldsymbol{\epsilon}_l)}{\sum_{k=0}^{L} |\lambda_1|^{2k}} \right| + 4\sqrt{N_x} \sum_{p=2}^{\frac{N_x+1}{2}} \left\{ |1-\nu_p| \frac{D_2}{(p-1)^{r+1}} + (|1-\nu_p| + 2\xi_p) \frac{D_3}{N_x^{r+1}} \right\} \left| \frac{\sum_{l=0}^{L} \overline{\lambda}_p^l \mathcal{F}_p(\boldsymbol{\epsilon}_l)}{\sum_{k=0}^{L} |\lambda_p|^{2k}} \right| \tag{4.65}$$

and D_1 is a constant independent of p, N_x and r and D_2 and D_3 are constants independent of p and N_x but dependent on r and ξ_p is defined as in (4.20).

Proof. Equation (4.58) gives that,

$$\tilde{\mathbf{x}}_0 - \mathbf{x}_a = (I - A_L)\tilde{\mathbf{x}}_0 - \boldsymbol{\rho}_L - \left[\sum_{r=0}^{L} (\Lambda^* \Lambda)^r \right]^{-1} \left[\sum_{l=0}^{L} (\Lambda^*)^l V^* \boldsymbol{\epsilon}_l \right]. \tag{4.66}$$

Then by taking norms and applying the triangle inequality,

$$\begin{aligned} \|\tilde{\mathbf{x}}_0 - \mathbf{x}_a\|_2^2 &\leq \sum_{p=1}^{N_x} \left\{ ||1-\nu_p||\mathcal{F}_p(\tilde{\mathbf{x}}_0)| - |\mathcal{F}_p(\boldsymbol{\rho}_L)|| + \left| \frac{\sum_{l=0}^{L} \overline{\lambda}_p \mathcal{F}_p(\boldsymbol{\epsilon}_l)}{\sum_{r=0}^{L} |\lambda_p|^{2r}} \right| \right\}^2 \\ &= \|(I - A_L)\tilde{\mathbf{x}}_0 - \boldsymbol{\rho}_L\|_2^2 + N_x z_0 \left(\left[\sum_{r=0}^{L} (\Lambda^* \Lambda)^r \right]^{-1} \sum_{l=0}^{L} (\Lambda^*)^l V^* \boldsymbol{\epsilon}_l \right) \\ &\quad + 2 \sum_{p=1}^{N_x} \{ |1-\nu_p||\mathcal{F}_p(\tilde{\mathbf{x}}_0)| + |\mathcal{F}_p(\boldsymbol{\rho}_L)| \} \left| \frac{\sum_{l=0}^{L} \overline{\lambda}_p \mathcal{F}_p(\boldsymbol{\epsilon}_l)}{\sum_{r=0}^{L} |\lambda_p|^{2r}} \right|. \end{aligned}$$

Using the result of lemma 4.4 together with (4.24) and (4.27) gives,

$$\|\tilde{\mathbf{x}}_0 - \mathbf{x}_a\|_2^2 \leq E_M + E_O + E_C,$$

where E_M, E_O and E_C are given by (4.63)-(4.65). □

The bound in (4.62) is formed from the equivalent bound in the absence of observation errors (E_M), together with the autocorrelation at lag 0 of the observation

error component of the analysis vector (E_O) and cross terms (E_C). We now analyse the behaviour of this bound in comparison to the l_2-norm of the error in the analysis vector, found through strong constraint 4D-Var data assimilation. The aim is to identify whether this bound can be used to characterise the behaviour of the error in the analysis vector. We do this by comparing the order of convergence to zero for the bound with the same for the l_2-norm of the error in the analysis vector, found through strong constraint 4D-Var data assimilation experiments.

The variables E_O and E_C are dependent on the random variables $\{\boldsymbol{\epsilon}_l\}_{l=0}^{L}$. However by the strong law of large numbers [80] if the experiments could be repeated, then as the number of experiments is increased, the sample means of E_O and E_C would tend toward their expected values. As a consequence, we consider the expected values of E_O and E_C;

$$\mathbb{E}\left[E_O\right] = \sigma_o^2 \sum_{p=1}^{N_x} \frac{1}{\sum_{k=0}^{L} |\lambda_p|^{2k}}, \text{ and } \mathbb{E}\left[E_C\right] = 0, \tag{4.67}$$

to identify the behaviour of the bound in (4.62) with respect to both N_x and L.

The expected value of E_C is zero whilst the expected value of E_O is dependent upon N_x, L, σ_o^2 and the numerically dissipative properties of the resolvable wavenumber components of the scheme. Hence the expected value is independent of both the regularity of the initial condition $u_0(x)$ and the dispersive properties of the resolvable wavenumber components of the finite difference scheme. A non-dissipative scheme leads to $\mathbb{E}\left[E_O\right] = \frac{\sigma_o^2 N_x}{L+1}$, so that the order of convergence for $\mathbb{E}\left[E_O\right]$ to zero with respect to N_x and L, is $\mathcal{O}(N_x)$ and $\mathcal{O}(L^{-1})$ respectively.

The order of convergence of the bound in (4.62) to zero, with respect to N_x or L, is determined by the dominant order of convergence possessed by either E_M or $\mathbb{E}[E_0]$. The orders of convergence to zero for E_M were analysed in Section 4.3. Table 4.9 displays the numerical orders of convergence to zero, with respect to N_x and L, for $\mathbb{E}\left[E_O\right]$. We see that the order of convergence with respect to N_x is the same for each scheme and we achieve the $\mathcal{O}(N_x)$ convergence we expected for the Preissman Box and MNIMC schemes. The order of convergence with respect to L is fairly small for the Upwind and Lax-Wendroff schemes, but we achieve the $\mathcal{O}(L^{-1})$ convergence we predicted for the Preissman Box and MNIMC schemes.

Variable	Upwind	Preissman Box	Lax-Wendroff	MNIMC
α	1.0000	1.0000	1.0000	1.0000
β	-3.3207×10^{-4}	-9.9719×10^{-1}	-2.0866×10^{-3}	-9.9719×10^{-1}

Table 4.9: *Numerical orders of convergence to zero, with respect to N_x and L, for $\mathbb{E}\left[E_O\right]$ in (4.67), given to 4dp, $\mathbb{E}\left[E_0\right] = \mathcal{O}(N_x^{\alpha} L^{\beta})$, with $h = 0.5$ and $\mu = 1$. The results for N_x and L were identified using fixed $L = 4$ ($\Delta t = \frac{1}{2N_x}$) and fixed $N_x = 3^7$ ($\Delta t = \frac{1}{2 \cdot 3^7}$), respectively. The results displayed here are the orders of convergence for the largest values of N_x and L considered, respectively.*

Initially consider the order of convergence of the bound in (4.62) with respect to

N_x. The results of Section 4.3 show that E_M remains constant or decays to zero, whilst Table 4.9 shows that $\mathbb{E}[E_O]$ increases, as N_x is increased. A similar property is seen for the order of convergence of each variable to zero, with respect to L. E_M increases, and $\mathbb{E}[E_O]$ decreases, as L is increased. Subsequently, the dominant order of convergence of (4.62) to zero, for both N_x and L, will be determined by the order of magnitude of the coefficients of each term.

Strong constraint 4D-Var numerical experiments were conducted using the set up detailed in Section 4.3. The results can be seen in Figures 4.9 and 4.10. The variance of the additive noise (σ_o^2) in the observations used to generate Figures 4.9 and 4.10, was chosen so that the Figures would best display the change in behaviour witnessed in all three schemes, as N_x and L are increased respectively.

Initially consider the numerical results for the Upwind, Preissman Box and Lax-Wendroff schemes. Figure 4.9 shows that initially the error in the analysis vector behaves according to E_M. Once a critical value of N_x has been reached for initial conditions with regularities greater than zero, the error then increases according to the behaviour exhibited by $\mathbb{E}[E_O]$ in Table 4.9. This provides a critical value for N_x at which the effect of both numerical model error due to finite difference approximations and observation errors on the accuracy of the analysis vector, is minimised. L and σ_o^2 form part of the coefficient when considering $\mathbb{E}[E_O]$ as a function of N_x. Increasing L or decreasing σ_o^2 will result in the critical value of N_x increasing, whilst decreasing L or increasing σ_o^2 will result in the critical value of N_x decreasing. The critical value for N_x shown in Figure 4.9 is between 3^4 and 3^5, depending on the chosen finite difference scheme.

Figure 4.10 shows a similar picture to that seen in Figure 4.9. However in this instance, the initial decrease in the error in the analysis vector corresponds $\mathbb{E}[E_O]$. As L is increased further, a critical value is reached where E_M becomes dominant over $\mathbb{E}[E_O]$, and the error begins to increase with L as described in Section 4.3.6. As with N_x, this critical value of L is determined by the coefficients of E_M and $\mathbb{E}[E_O]$. Decreasing either N_x or σ_o^2 will result in the critical value of L decreasing, whilst increasing either N_x or σ_o^2 will result in the critical value of L increasing.

Rabier et al. [4] found that their incremental 4D-Var experiments improved upon the results of their 3D-Var experiments, for assimilation windows of 6 or 12 hours. However, this was not the case when a 24 hour assimilation window was considered. This was explained in part by the use of the tangent-linear and adjoint models, which contain errors as they are approximations of the fully non-linear models [4]. Increasing L in our experiments, extends the length of the assimilation window. Our results also found that this leads to an increase in the error in our analysis vector, once past a critical value of L, despite our problem not containing any of the error due to approximations of the fully non-linear models. However, our experiments did not utilise incremental 4D-Var as we were considering a linear problem, but our problem could be viewed as implementing the inner loop of incremental strong constraint 4D-Var, where M is

our tangent-linear model. Our results show how errors in the implementation of the tangent-linear and adjoint models could possibly create errors in the analysis increment. This form of error is could also be affecting the results of Rabier et al. [4].

When considering the orders of convergence with respect to either N_x or L, reducing σ_o^2 corresponds to reducing the error in the observations. As a result, it is not surprising that reducing σ_o^2 results in E_M becoming the dominant order of convergence, with respect to both N_x or L.

If we now consider the MNIMC scheme, Figure 4.9 shows that the l_2-norm of the error in the analysis vector grows as N_x increases, with the order of convergence of $\mathbb{E}[E_0]$ with respect to N_x. This is due to the numerical model error introduced into the analysis vector by the MNIMC is quite small in comparison to observation errors, as the only source is aliasing errors. This reinforces our conclusions for the Upwind, Preissman Box and Lax-Wendroff schemes, that the increase in the error present, is due to the effects of observation errors. Figure 4.10 shows that as we increases L, the error in the analysis vector initially decays for the MNIMC scheme, similarly to the Upwind, Preissman Box and Lax-Wendroff schemes, due to observation errors. However, it is unclear without further numerical experiments where L is increased further, whether a critical value of L will be reached where the error in the analysis vector begins to increase. It may be that due to the small size of the numerical model error in the analysis vector, that we do not see this increase.

This analysis suggests that (4.62) is an appropriate bound to demonstrate the order of convergence for the error in the analysis vector. As a result, given a fixed value for σ_o^2 and either N_x or L, it is possible to choose a value for L and N_x respectively, that minimises the error in the analysis vector due to numerical model error and observation errors. The result for N_x in some way works towards answering the question posed by Akella and Navon [47], as to whether increasing the number of discretisation points would continue to decrease the effects of discretisation errors on the results of strong constraint 4D-Var data assimilation. In this instance, we have shown that when considering numerical model due to finite difference approximations and observation errors in strong constraint 4D-Var, increasing the number of discretisation points when considering full sets of observations, past the optimal value of N_x would result in an increase in the error in the analysis vector.

Furbish et al. [52] investigated how the number of discretisation points of the numerical model, the density of observations and interpolation errors affect the accuracy of the forecast from the analysis vector, using a shallow water finite volume model. Perfect observations were considered, but background error and interpolation errors played a role in the problem. It was found that increasing the density of observations whilst the numerical model possessed a given number of discretisation points, did not improve the results of the forecast from the analysis vector. However increasing the number of discretisation points did result in an improved forecast when the density of observations was increased. This revealed that the resolution of the model needs to be

sufficiently large for an increase in the observations to improve the forecast [52]. We have not examined the effect of the number of discretisation points and observations as independent entities to avoid interpolation errors. By this we mean that increasing the number of discretisation points does, as a by product of the way we take observations, increase the number of observations. We also do not include a background term or any interpolation errors. Despite these differences, our results seem to be similar to Furbish et al. [52]. The critical value of L at which the error in the analysis vector increases, increases as N_x increases. These means that there is a larger range of values to choose for L where a decrease in the error would be seen in the analysis vector for our problem. The forecast would then be expected to be more accurate. If N_x is too small, this range of values will decrease. More research is required as Furbish et al. [52] incorporates many other forms of error, but it is encouraging that the results from our experiments on the 1D linear advection problem may also be being seen for a shallow water problem.

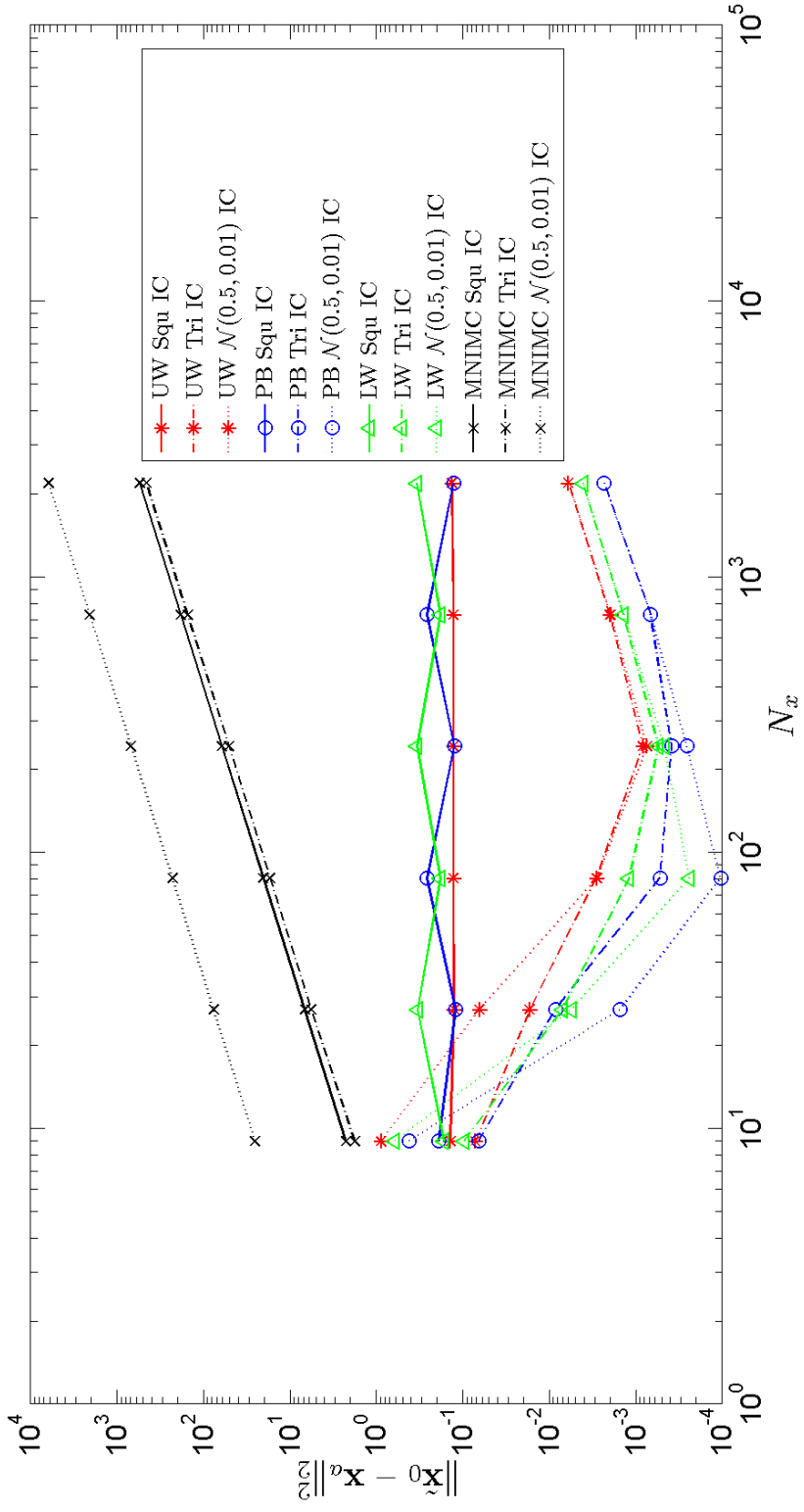


Figure 4.9: *The square of the l_2-norm of the error in the analysis vector, calculated through strong constraint 4D-Var data assimilation numerical experiments, under the influence of errors introduced by finite difference approximations in the forward model and observation errors. The observations are Gaussian random variables with mean zero and variance σ_o^2. The results were generated using the Upwind (UW), Preissman Box (PB), Lax-Wendroff (LW) and MNIMC schemes as the forward models for solving the 1D linear advection problem in (3.1), using $h = 0.5$, $\mu = 1$, $L = 4$ and $\sigma_o^2 = 5 \times 10^{-6}$, where $N_x = 3^\gamma$ for $\gamma = 2, \ldots, 7$ ($\Delta t = \frac{1}{2N_x}$). The functions considered for $u_0(x)$ in these experiments are defined in Section 4.3, where 'squ IC', 'tri IC', $\mathcal{N}(0.5, 0.01)$ IC' denote the 1D square, the triangular and 1D Gaussian functions respectively. The results are plotted using logarithmic scales to demonstrate the order of convergence of the error to zero, with respect to N_x.*

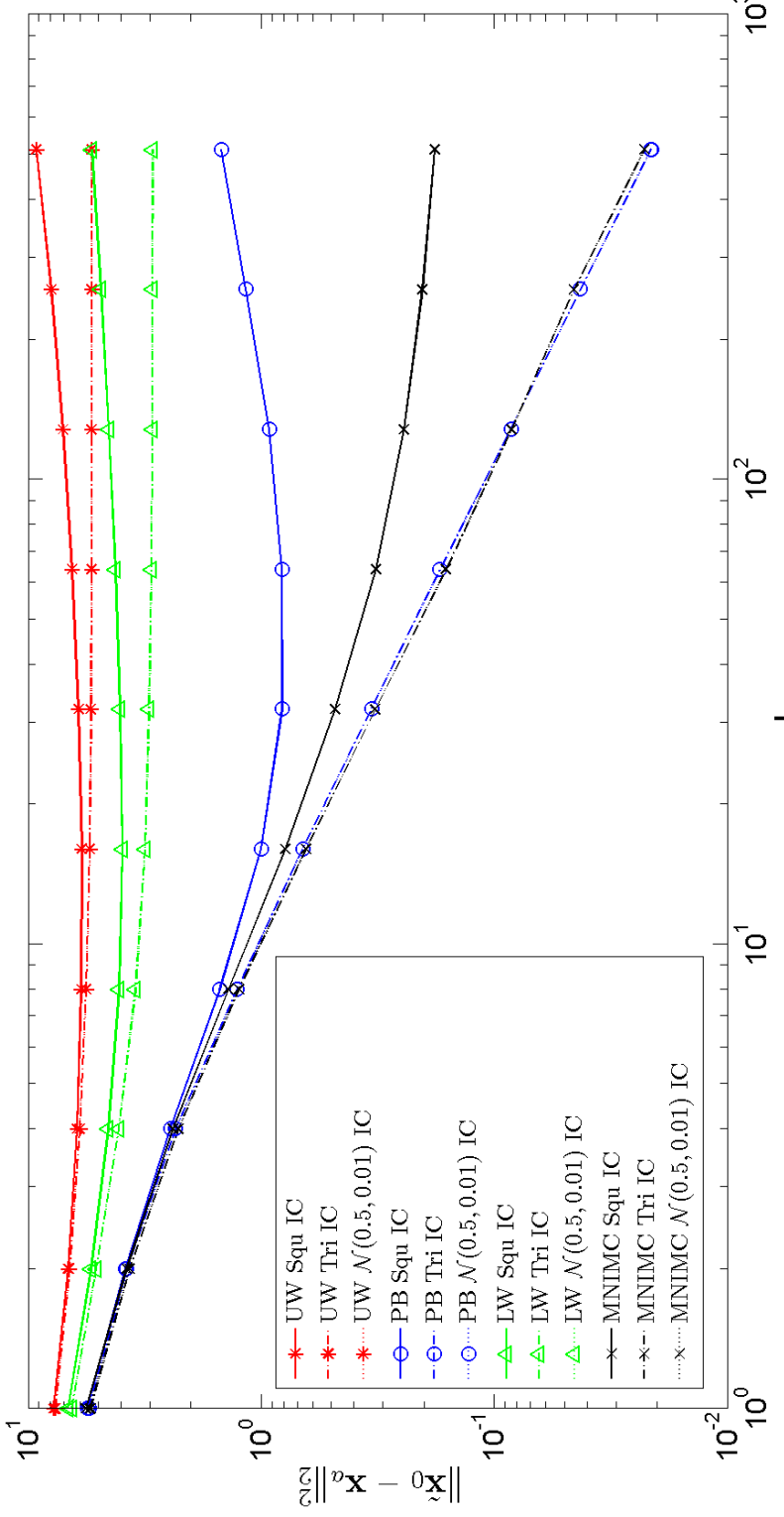


Figure 4.10: *The square of the l_2-norm of the error in the analysis vector, calculated through strong constraint 4D-Var data assimilation numerical experiments, under the influence of errors introduced by finite difference approximations in the forward model and observation errors. The observations are Gaussian random variables with mean zero and variance σ_o^2. The results were generated using the Upwind (UW), Preissman Box (PB), Lax-Wendroff (LW) and MNIMC schemes as the forward models for solving the 1D linear advection problem in (3.1), using $h = 0.5$, $\mu = 1$, $N_x = 3^7$ and $\sigma_o^2 = 5 \times 10^{-3}$, where $L = 2^\delta$ for $\delta = 0, \ldots, 9$ ($\Delta t = \frac{1}{2 \cdot 3^7}$). The functions considered for $u_0(x)$ in these experiments are defined in Section 4.3, where 'squ IC', 'tri IC', $\mathcal{N}(0.5, 0.01)$ IC' denote the 1D square, the triangular and 1D Gaussian functions respectively. The results are plotted using logarithmic scales to demonstrate the order of convergence of the error to zero, with respect to L.*

4.5 Relevance of the results to weak constraint 4D-Var

The work presented in Section 3.10 can be used to choose the relevant time dependent function for deterministic errors in weak constraint 4D-Var data assimilation of the considered 1D linear advection problem. The model formulation in Equation (2.5) of Section 2.5 provides a continuous formulation for the physical system in weak constraint 4D-Var data assimilation. The 1D linear advection problem is considered as the physical system for the strong constraint 4D-Var problem in Section 2.3. This means that the model equations are perfect for our considered problem, so the weak constraint model equations for our considered problem cannot initially be formulated as in (2.5). The model error we are analysing is numerical model error arising from the use of finite differences to approximate derivatives. Therefore these errors arise in the discretised model, so their correction is initially formulated in the discretisation of the model equations. Consider the following formulation for the numerical model, for use with weak constraint 4D-Var,

$$\begin{aligned} &\mathbf{x}_l = M^l\mathbf{x}_0 + \boldsymbol{\eta}_l, && \text{for } l \in \mathbb{N}_0, \\ &\boldsymbol{\eta}_l = \Phi_{l-1}\boldsymbol{\eta}_{l-1}, && \text{for } l \in \mathbb{N}\setminus\{1\}, \\ &\mathbf{x}_0 = \mathbf{x}(0), \;\; \boldsymbol{\eta}_0 = \mathbf{0}, \;\; \boldsymbol{\eta}_1 = (\tilde{M} - M)\mathbf{x}_0. \end{aligned} \tag{4.68}$$

Here $\tilde{M}$ is the matrix implementing the MNIMC scheme, $\Phi_l \in \mathbb{R}^{N_x \times N_x}$ denotes the deterministic error evolution matrix and $\boldsymbol{\eta}_l \in \mathbb{R}^{N_x}$ denotes the deterministic model error correction term. Comparing Equation (4.68) with Equation (2.8), we can see that in the case of Equation (4.68) $\mathcal{T}_l = I_{N_x}$ the $N_x \times N_x$ identity matrix and $\Phi(\mathbf{x}_l, \boldsymbol{\eta}_l) = \Phi_l$ the deterministic error evolution matrix. The choice of $\boldsymbol{\eta}_0$ and $\boldsymbol{\eta}_1$ seen in Equation (4.68) will be justified in the following. Examining the error in the lth numerical model state of the alternative model in (4.68) results in,

$$\mathbf{e}_l := \tilde{\mathbf{y}}_l - \mathbf{x}_l = \begin{cases} (\tilde{M}^l - M^l)\mathbf{x}_0 + \mathbf{r}_l - \boldsymbol{\eta}_l, & \text{for } l \in \mathbb{N}, \\ -\boldsymbol{\eta}_0, & \text{for } l = 0. \end{cases} \tag{4.69}$$

The vector $\mathbf{r}_l$ is defined as in Equation (3.53). Since $\boldsymbol{\eta}_l$ must correct for the deterministic error in $\mathbf{x}_l$, ie: $(\tilde{M}^l - M^l)\mathbf{x}_0$, using (4.69), we define,

$$\boldsymbol{\eta}_l := \begin{cases} (\tilde{M}^l - M^l)\mathbf{x}_0, & \text{for } l \in \mathbb{N}, \\ \mathbf{0}, & \text{for } l = 0, \end{cases} \tag{4.70}$$

hence determining $\boldsymbol{\eta}_l$ in (4.68). Then the deterministic error evolution matrix is defined as,

$$\Phi_l := \left(\tilde{M}^l - M^l\right)\left(\tilde{M}^{l-1} - M^{l-1}\right)^{-1} \quad \text{for } l \in \mathbb{N}, \tag{4.71}$$

provided $(\tilde{M}^{l-1} - M^{l-1})$ is an invertible matrix. The resulting cost function is then,

$$J(\mathbf{x}_0, \boldsymbol{\eta}_0) = \sum_{l=0}^{L} (\mathbf{y}_l - M^l \mathbf{x}_0)^T (\mathbf{y}_l - M^l \mathbf{x}_0) + (\boldsymbol{\eta}_b - \boldsymbol{\eta}_0)^T Q^{-1} (\boldsymbol{\eta}_b - \boldsymbol{\eta}_0), \tag{4.72}$$

as given in Akella and Navon [47], where $Q \in \mathbb{R}^{N_x \times N_x}$ is the error covariance matrix for the initial estimate of $\boldsymbol{\eta}_0$ denoted by $\boldsymbol{\eta}_b \in \mathbb{R}^{N_x}$. Here $\boldsymbol{\eta}_0$ is included in Equation (4.72) to demonstrate the role of $\boldsymbol{\eta}_0$ despite it possessing a value of $\mathbf{0}$ in this instance.

However, the deterministic error evolution matrix only accounts for the errors present in the resolvable wavenumber components of each numerical model state. In reality there are also aliasing errors present in the problem, denoted by $\mathbf{r}_l$ in (4.69), but quantifying these errors is not possible unless we know the true solution. Griffith and Nichols [45] and Akella and Navon [47] both propose augmenting the model error in the lth state with a random component, $\boldsymbol{\zeta}_l \in \mathbb{R}^{N_x}$. This idea could be applied to this problem to account for aliasing errors.

However, the new cost function would also need to be minimised with respect to all of these random errors, increasing the computational time and cost to find the optimal solution. In the case of the 1D linear advection problem, equation (3.55) describes the shifted $b\Delta t$-periodic nature of the aliasing errors $\mathbf{r}_l$, introduced by the MNIMC scheme. This property means that only the first $b-1$ random errors need be used as control variables in the minimisation. Hence (4.68) can be augmented using (3.55) to attempt to correct for aliasing errors using random variables,

$$\begin{aligned} &\mathbf{x}_l = M^l \mathbf{x}_0 + \boldsymbol{\eta}_l, && \text{for } l \in \mathbb{N}_0, \\ &\boldsymbol{\eta}_l = \Phi_{l-1} \boldsymbol{\eta}_{l-1} - \tilde{M}^{l-[l]_b} \boldsymbol{\zeta}_{[l]_b}, && \text{for } l \in \mathbb{N} \setminus \{1\}, \\ &\boldsymbol{\eta}_0 = \mathbf{0}, \;\; \boldsymbol{\eta}_1 = \left(\tilde{M} - M\right) \mathbf{x}_0. \end{aligned} \tag{4.73}$$

The resulting cost function is,

$$\begin{aligned} J(\mathbf{x}_0, \boldsymbol{\eta}_0, \boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_{b-1}) \;\; = \;\; & \sum_{l=0}^{L} (\mathbf{y}_l - M^l \mathbf{x}_0)^T (\mathbf{y}_l - M^l \mathbf{x}_0) \\ & + (\boldsymbol{\eta}_b - \boldsymbol{\eta}_0)^T Q^{-1} (\boldsymbol{\eta}_b - \boldsymbol{\eta}_0) + \sum_{l=1}^{b-1} \boldsymbol{\zeta}_l G_l^{-1} \boldsymbol{\zeta}_l, \end{aligned} \tag{4.74}$$

as demonstrated in Griffith and Nichols [45] for random errors. Here we neglect $\boldsymbol{\zeta}_0$ as $\mathbf{r}_0 := \mathbf{0}$. The matrix $G_l \times \mathbb{R}^{N_x \times N_x}$ is the error covariance matrix for the random error $\boldsymbol{\zeta}_l$ for each l.

We now wish to numerically implement this weak constraint 4D-Var for our considered schemes. Consider the deterministic error evolution matrix, defined in (4.71), for the Upwind, Preissman Box and Lax-Wendroff schemes. Using the eigenvalue decompositions of the matrix M implementing one of these schemes and the matrix $\tilde{M}$

implementing the MNIMC scheme, we obtain,

$$\Phi_l = V\left(\tilde{\Lambda}^l - \Lambda^l\right)\left(\tilde{\Lambda}^{l-1} - \Lambda^{l-1}\right)^{-1} V^*. \tag{4.75}$$

The matrix $\tilde{\Lambda}^{l-1} - \Lambda^{l-1}$ is not invertible if this matrix has a zero eigenvalue. This occurs when an eigenvalue of the matrix M is non-dissipative and non-dispersive with respect to a resolvable wavenumber component, ie: $\exists p$ such that $\lambda_p = \tilde{\lambda}_p$ for $p = 1, \ldots, N_x$. In this instance, the corresponding eigenvalue of $\tilde{\Lambda}^l - \Lambda^l$ is also zero. The case of a non-dissipative and non-dispersive eigenvalue of the matrix M, arises in λ_1 for the Upwind, Preissman Box and Lax-Wendroff schemes and for all eigenvalues when the MNIMC scheme is considered. Therefore, for the implementation of this weak constraint 4D-Var in these cases, we need an alternate formulation for Φ_l.

Suppose the eigenvalue λ_p of M for some $p = 1, \ldots, N_x$, is non-dissipative and non-dispersive, so it correctly propagates its corresponding resolvable wavenumber component. This results in the deterministic error in the corresponding resolvable wavenumber component of the numerical solution generated by M, being equal to zero, using our definition of deterministic error in (4.69). Therefore, we require that $\mathcal{F}_p(\boldsymbol{\eta}_l) = 0$ for all $l \in \mathbb{N}_0$. It then makes sense to define the pth eigenvalue of Φ_l to be zero, for all $l \in \mathbb{N}_0$. Therefore we now define the deterministic error evolution matrix by $\Phi_l = V\Upsilon_l V^*$, where $\Upsilon_l \in \mathbb{C}^{N_x \times N_x}$ is defined as the diagonal matrix of eigenvalues of Φ_l such that,

$$\{\Upsilon_l\}_{p,q} = \begin{cases} \frac{\tilde{\lambda}_p^l - \lambda_p^l}{\tilde{\lambda}_p^{l-1} - \lambda_p^{l-1}}\delta_{p,q}, & \text{for } \lambda_p \neq \tilde{\lambda}_p, \\ 0, & \text{for } \lambda_p = \tilde{\lambda}_p, \end{cases} \tag{4.76}$$

for all $l \in \mathbb{N}$. The formulation for the weak constraint 4D-Var cost function presented in this Section, needs to be tested to identify how it affects the accuracy of the analysis vector for the 1D linear advection problem, when finite difference approximations are the only form of error. If this test is successful, then the original cost function can also be augmented in this way and this formulation of weak constraint 4D-Var data assimilation, tested further other forms of error associated with strong constraint 4D-Var data assimilation.

4.6 Summary

This Chapter aimed to construct a bound for the l_2-norm of the error in the analysis vector, solely under the effects of numerical model error introduced by finite difference approximations in the forward model. A spectral approach was found to be effective constructing a bound that represents the worst case error in the analysis vector. This bound was dependent on the regularity (r) of the true initial condition $u_0(x)$ we wish to recover, the numerically dissipative and dispersive properties of the resolvable wavenumber components of the finite difference scheme used as the forward model, the

number of discretisation points (N_x) when considering full sets of observations and the number of sets of observations in the assimilation window (L). Here the regularity of a function represents the smoothness of the function.

The bound was re-written as a sum of six terms S_1 to S_6, multiplied by coefficients independent of N_x and L. Each of the terms S_1 to S_6 is a summation dependent on r, N_x, L and the numerically dissipative and dispersive properties with respect to the resolvable wavenumber components of the finite difference scheme. The order of convergence of S_1 to S_6 to zero was identified numerically with respect to both N_x and L, for various values of r, using the Upwind, Preissman Box, Lax-Wendroff and MNIMC schemes. These results were compared against the order of convergence to zero with respect to N_x and L, for the l_2-norm of the error in the analysis vector, found through strong constraint 4D-Var data assimilation numerical experiments, for selected regularity true initial conditions and the same finite difference schemes.

Comparing these results we find that the bound for the l_2-norm of the error in the analysis vector, is suitable for characterising the behaviour of the l_2-norm of the error in the analysis vector introduced by finite difference schemes, for the Upwind, Preissman Box and Lax-Wendroff schemes. It was sufficient to analyse summation S_1 multiplied by its corresponding coefficient from the bound to characterise this behaviour, and was found to possibly form a tighter bound for the error introduced by these schemes, showing that numerically dissipative and dispersive effects on resolvable wavenumber components has a larger effect than aliasing errors. S_4 is the only component of the bound on the l_2-norm of the error in the analysis vector for the MNIMC scheme. This summation was found to characterise the behaviour of the error for small regularities. More work needs to be conducted to understand why.

Asymptotic expansions were created for each summations comprising the bound for the Upwind and MNIMC schemes. The expansions for the Upwind scheme were able to analytically demonstrate the numerical behaviour of each summation with respect to N_x but not L. The asymptotic expansion of S_4 for the MNIMC scheme was only able to represent the numerical behaviour of S_4 with respect to N_x and L for small regularity true initial conditions.

The order of convergence to zero for the l_2-norm of the error in the analysis vector with respect to N_x and L, was found through strong constraint 4D-Var numerical experiments, for each considered scheme and true initial conditions with varying regularities. Analysing these results found that the error in the analysis vector remained constant as N_x was increased, for discontinuous true initial conditions, for all considered schemes. The error in the analysis vector decayed for smoother true initial conditions, at a constant rate determined by r, as N_x increased. Once a critical regularity was reached, the rate no longer increased with r. A decay in the error with respect to the number of discretisation points is something you might expect, but the constant error for discontinuities was a surprising result and may be a result of Gibb's phenomenon. The error in the analysis vector increased as the number of sets of observations in the

assimilation window increased. This was an unexpected result as you might expect that more information would increase the accuracy of the analysis vector. This result agrees with the numerical results of Griffith [81]. Increasing the number of observations in the assimilation window means that the estimated initial condition needs to form a good fit to more sets of observations [82], causing the error in the analysis vector to increase [81].

Re-introducing observation errors to the problem and performing a similar analysis revealed that when considering smooth true initial conditions ($r \in \mathbb{N}$), there is a critical number of observations when considering full sets of observations (a critical density of observations), where strong constraint 4D-Var data assimilation can be performed, that minimises the impact of errors due to finite difference approximations and observation errors for the Upwind, Preissman Box and Lax-Wendroff schemes. More research is required on this result as in reality there are many other forms of error which also affect the accuracy of the analysis vector. There are also many other meteorological relevant PDE systems which need to be investigated to see if such a property exists. Considering the MNIMC in this problem, we find that observation errors dominate the aliasing errors introduced by the scheme. As a result the error is always increasing due to observation errors and is larger than for any of the other considered schemes for the same number of discretisation points and regularity of the true initial condition. This indicates that a numerically dissipative and/or dispersive finite difference scheme with respect to the resolvable wavenumber components of the numerical solution, could be an advantage for minimising the affects of observation errors on the accuracy of the analysis vector.

The contribution from white noise observations to the analysis vector was also analysed. This found that the effects of numerical dissipation in the resolvable wavenumber components of a scheme, can lead to correlations in the white noise contribution to the analysis vector. This could possibly lead to artifacts in the analysis vector which could potentially be introduced into any forecast formed from the analysis vector. Despite numerically non-dissipative and dispersive schemes with respect to the resolvable wavenumber components of the numerical solution (Preissman Box scheme) introducing destructive interference, resulting in a loss of information in the analysis vector, such a scheme would have a critical value of N_x when considering full sets of observations, for performing strong constraint 4D-Var data assimilation for smooth true initial conditions and would prevent artifacts from being introduced to the analysis vector. This could mean that if the critical value of N_x is fairly small, then this may be the best choice of scheme. This hypothesis needs to be tested by re-introducing other forms of error to the problem.

This Chapter was also able to construct a deterministic model error operator for use in weak constraint 4D-Var data assimilation of the 1D linear advection problem. The formulation of weak-constraint 4D-Var data assimilation using a deterministic model error operator was suggested by Akella and Navon [47]. The model equations were

also augmented with a random term, as suggested by Griffith and Nichols [45]. In this instance, the random component was used to account for the aliasing errors introduced by the MNIMC scheme. The shifted periodic nature of this aliasing error reduced the number of random variables to be estimated. The model formulation proposed for use in weak constraint 4D-Var data assimilation of the 1D linear advection problem, needs to be tested to determine its ability to reduce the impact of errors due to finite difference approximation, on the analysis vector. This is left as future work.

It should be noted that the results from Chapters 3 and 4 were generated for chosen representative finite difference schemes; the Upwind, Preissman Box, Lax-Wendroff and MNIMC schemes. However the theory is applicable to any finite difference scheme that can be implemented to solve the 1D linear advection equation in a similar way. The results presented in this Chapter are for a linear problem. Most practical applications of 4D-Var are for non-linear problems, so it is important to address the question of how these results relate to non-linear problems in 4D-Var. Pfeffer et al. [83] performed a similar linear analysis of the numerically dissipative and dispersive properties of the Matsuno and Leapfrog schemes before using them to solve a non-linear problem in the NASA-GLAS climate model. Their results showed that some of the properties of the schemes found through this analysis could be seen in the results of the non-linear problem. This leads us to believe that the results of the linear analysis presented here, are to some extent, directly relevant to non-linear problems. Future work would be to compare the results of this Chapter with those of a non-linear problem solved with these schemes, as in Pfeffer et al. [83]. It would also be interesting to compare the results from this Chapter with those from incremental 4D-Var, where the 1D linear advection problem forms the linearised physical system. In the next Chapter the work conducted in this Chapter, analysing the behaviour of the l_2-norm of the error in the analysis vector found through strong constraint 4D-Var data assimilation, is extended to a similar 2D linear advection problem.

CHAPTER 5

The 2D Linear Advection Problem

In this chapter, we continue our investigation of the effects of numerical model error, introduced by finite difference approximations in the forward model, on the analysis vector obtained through strong constraint 4D-Var data assimilation. We now consider the 2D linear advection equation, together with circulant boundary conditions and initial condition, as our physical system of interest. The main advantage of considering this problem is that it allows us to extend our results from Chapters 3 and 4 to the equivalent 2D problem, making use of our understanding of the 1D problem. It also allows us to investigate any difficulties which arise from applying the same analysis techniques to a 2D problem.

Chapter 4 revealed some interesting results on the effects of numerical dissipation and dispersion, the smoothness of the initial condition, the number of discretisation points when considering full sets of observations and the number of sets of observations in the assimilation window, on the behaviour l_2-norm of the error in the analysis vector. Extending these results to the 2D problem is of importance, due to the extra computational expenses involved in computing the solution to a 2D problem. If an optimal number of discretisation points when considering full sets of observations or number of sets of observations in the assimilation window could be chosen to minimise the effects of errors on the analysis vector, then this would justify an increase or decrease in the computational resources required.

In order to accomplish our aim of producing results similar to those in Chapter 4, we require the construction of a bound on the l_2-norm of the error in the analysis vector and numerical results to compare it against. As we are considering the same strong constraint 4D-Var data assimilation problem set out in Section 2.3 and wish to perform the same analysis as in Chapter 4, we must begin by formulating the analysis vector similarly to Chapter 3. Therefore we begin with the same analysis as in Chapter 3, using 2D Fourier series to investigate aliasing and to define numerical dissipation and dispersion. We choose to investigate the Upwind and Crank-Nicolson finite difference

schemes due to their numerically dissipative and dispersive properties. Perfect observations for the problem can be constructed algebraically using the technique derived in Chapter 3, using the MNIMC scheme for the 2D problem.

In order to construct a bound on the l_2-norm of the error in the analysis vector using our spectral approach, we need bounds on the 2D Fourier coefficients and the error in the coefficient found through the 2D DFT, compared to the corresponding 2D Fourier coefficient. We derive these bounds for multiplicatively separable functions and discuss the difficulties that arise for functions without this property. Once these bounds have been derived, a bound on the l_2-norm of the error in the analysis vector can be constructed and numerical results computed for comparison to motivate the future analysis of the bound.

5.1 The physical system

Consider the 2D linear advection equation for the function, $u : \mathbb{R} \times \mathbb{R} \times [0, \infty) \to \mathbb{R}$, $(x, y, t) \mapsto u(x, y, t)$, together with circulant boundary conditions and initial condition $u_0 : [0, 1) \times [0, 1) \to \mathbb{R}$, $(x, y) \mapsto u_0(x, y)$,

$$\begin{aligned} &u_t + \mu_1 u_x + \mu_2 u_y = 0, && x \in [0,1), \;\; y \in [0,1), \;\; t > 0, \\ &u(x,y,t) = u(x+1,y,t), && x \in \mathbb{R}, \;\; y \in \mathbb{R}, \;\; t \geq 0, \\ &u(x,y,t) = u(x,y+1,t), && x \in \mathbb{R}, \;\; y \in \mathbb{R}, \;\; t \geq 0, \\ &u(x,y,0) = u_0(x,y), && x \in [0,1), \;\; y \in [0,1). \end{aligned} \tag{5.1}$$

This is a linear, hyperbolic, two-dimensional PDE problem. Similarly to the 1D linear advection problem, here the *wave speeds* in the x- and y-directions are given by μ_1 and μ_2 respectively, $\mu_1, \mu_2 \in \mathbb{R}$ [13]. The 2D linear advection equation is also considered in the context of data assimilation by Vukićević et al. [55]. It is important to note that the scalar y is the spatial dimension, whilst the vectors $\{\mathbf{y}_l\}_{l=0}^{L}$ in Section 2.3 denote the sets of observations of the physical system.

The solution to this problem, $u(x, y, t) = u(x - \mu_1 t, y - \mu_2 t, 0) = u_0([x - \mu_1 t]_1, [y - \mu_2 t]_1)$ [13], behaves similarly to that of the 1D linear advection problem. Here $[\cdot]_1$ denotes modulo one. The shape of the initial condition is preserved over time and propagates in the x- and y-directions with speeds μ_1 and μ_2 respectively.

Problem (5.1) can also be solved numerically using a finite difference scheme, as the forward model. As with the analysis of the 1D linear advection problem, the numerical model error introduced into the strong constraint 4D-Var data assimilation problem by these schemes, will be analysed in terms of numerical dissipation and numerical dispersion. The analysis of the 1D linear advection problem required the use of Fourier series. As we are following the same analysis, we begin by reviewing two-dimensional Fourier series.

5.2 2D Fourier series

Consider a general function $f : \mathbb{R} \times \mathbb{R} \times [0, \infty) \to \mathbb{R}$, $(x, y, t) \mapsto f(x, y, t)$, such that $f(x+T_1, y+T_2) = f(x, t)$ for all t, for finite $T_1, T_2 \in \mathbb{R}^+$. This makes $f(x, y)$ T_1-periodic in x and T_2-periodic in y. The exponential form of the 2D Fourier series for $f(x, y, t)$ is given by $S : \mathbb{R} \times \mathbb{R} \times [0, \infty) \to \mathbb{R}$, such that $(x, t) \mapsto S(x, y, t)$ [60],

$$f(x, y, t) \sim S(x, y, t) = \sum_{p=-\infty}^{\infty} \sum_{q=-\infty}^{\infty} a_{p,q}(t) e^{\frac{2\pi i p x}{T_1}} e^{\frac{2\pi i q y}{T_2}}, \tag{5.2}$$

where,

$$a_{p,q}(t) = \frac{1}{T_1 T_2} \int_0^{T_1} \int_0^{T_2} f(x, y, t) e^{\frac{-2\pi i p x}{T_1}} e^{\frac{-2\pi i q y}{T_2}} \mathrm{d}y \, \mathrm{d}x. \tag{5.3}$$

As for the 1D Fourier series, this constructs an infinite sum representation of $f(x, y, t)$ from the superposition of 2D Fourier basis functions constructed from 1D Fourier basis functions in the x-direction, $e^{\frac{2\pi i p x}{T_1}}$ and in the y-direction, $e^{\frac{2\pi i q y}{T_2}}$. These are orthonormal basis functions in the x- and y-directions respectively [60] and together make orthonormal basis functions for $\mathbb{R}^2$. The 2D Fourier basis function $e^{\frac{2\pi i p x}{T_1}} e^{\frac{2\pi i q y}{T_2}}$ has wavelength $\frac{T_1}{p}$ and $\frac{T_2}{q}$ in the x- and y-directions respectively. Consequently, this 2D Fourier basis function has wavenumber $\frac{p}{T_1}$ and wavenumber $\frac{q}{T_2}$ in the x- and y-directions respectively. We term the (p, q)th 2D Fourier basis function multiplied by its corresponding Fourier coefficient in the considered 2D Fourier series, the (p, q)th wavenumber component of the 2D Fourier series, $p, q \in \mathbb{Z}$.

The coefficients $a_{p,q}(t)$ determine the contribution of each 2D Fourier basis function in the construction of $f(x, y, t)$. Many of the properties of the 1D Fourier coefficients, extend to the 2D Fourier coefficients. As $f(x, y, t)$ is a real-valued function and the (p, q)th and $(-p, -q)$th Fourier basis functions are complex conjugates, the coefficients have the property that $\overline{a_{p,q}(t)} = a_{-p,-q}(t)$ for all t. As a result, for all $p, q \in \mathbb{Z}$, not both zero, the contribution of the (p, q)th wavenumber component can be summed with the equivalent for the $(-p, -q)$th wavenumber component, to create a real valued function similarly to Equation (3.4). The wavenumber component for $p = q = 0$ and its coefficient $a_{0,0}(t)$ are both real functions. It is these quantities we will refer to as real wavenumber components of the 2D Fourier series.

The function $u(x, y, t)$ in problem (5.1) is 1-periodic in the x- and y-directions, so $T_1 = T_2 = 1$. Then the Fourier series of $u(x, t)$ is given by,

$$u(x, y, t) \sim \sum_{p=-\infty}^{\infty} \sum_{q=-\infty}^{\infty} b_{p,q}(t) e^{2\pi i p x} e^{2\pi i q y}, \tag{5.4}$$

$$\text{where } b_{p,q}(t) = \int_0^1 \int_0^1 u(x, y, t) e^{-2\pi i p x} e^{-2\pi i q y} \mathrm{d}y \, \mathrm{d}x.$$

We also define the Fourier series for $u_0(x, y)$,

$$u_0(x,y) \sim \sum_{p=-\infty}^{\infty} \sum_{q=-\infty}^{\infty} c_{p,q} e^{2\pi i p x} e^{2\pi i q y}, \tag{5.5}$$
$$\text{where } c_{p,q} = \int_0^1 \int_0^1 u_0(x,y) e^{-2\pi i p x} e^{-2\pi i q y} \mathrm{d}y \, \mathrm{d}x.$$

As the function $u_0(x, y)$ is only defined on $[0, 1) \times [0, 1)$, the Fourier series for $u_0(x, y)$ forms a Fourier series for the 1-periodic extension of $u_0(x, y)$, in both the x- and y-directions. The function $u(x, y, 0)$ is defined as the 1-periodic extension of $u_0(x, y)$, so they are represented by the same Fourier series. Hence, $b_{p,q}(0) := c_{p,q}$ for all $p, q \in \mathbb{Z}$.

To our knowledge, the literature surrounding two-dimensional Fourier series does not contain a Theorem similar to Theorem 3.1, setting out sufficient conditions for the convergence of two-dimensional Fourier series. Consequently the Lemmas in this Chapter which require a 2D Fourier series to be convergent, will not state the conditions the functions should satisfy to possess a convergent Fourier series. Instead, the considered functions will be defined as possessing a convergent Fourier series. As the 2D Fourier basis functions are constructed from the 1D Fourier series basis functions, it is imaginable that the only types of discontinuity allowed to be present in a two-dimensional function with a convergent Fourier series, are jump discontinuities. We will work under this assumption in this Chapter. Jump discontinuities can arise in the form of point discontinuities or if a function is piecewise continuous, a line of discontinuities forming a boundary of a continuous piece of the domain. It is possible for a point of discontinuity along such a line, to be discontinuous in the x-direction, y-direction or both. We would expect for Gibb's phenomenon to occur at these points of discontinuity, as the truncated Fourier series converges to the Fourier series for the function, similarly to 1D Fourier series. In the following Section we will consider the Upwind and Crank-Nicolson schemes for solving the 2D linear advection problem in (5.1).

5.3 Finite difference scheme formulation in 2D

Finite difference schemes used to numerically solve problem (5.1) are designed in the same way as they were for problem (3.1) in Section 3.1. There are many finite difference schemes available to solve problem (5.1), but we choose to consider the Upwind and Crank-Nicolson schemes. As before, these are used as 'representative schemes' chosen due to their numerically dissipative and dispersive properties.

In order to define a finite difference scheme over the domain $[0, 1] \times [0, 1]$, we require the following assumptions.

Assumption 5.1. *Divide the domain* $[0,1]\times[0,1]$ *into* N_x+1 *and* N_y+1 *equally spaced mesh points in the* x*-direction and* y*-direction respectively,* $N_x, N_y \in \mathbb{N}$*. This gives a grid spacing of* $\Delta x = \frac{1}{N_x}$ *and grid points* $x_j = j\Delta x$ *for* $j = 0,\dots,N_x$*, in the* x*-direction and* $\Delta y = \frac{1}{N_y}$ *and grid points* $y_k = k\Delta y$ *for* $k = 0,\dots,N_y$*, in the* y*-direction. Again we define the time step* $\Delta t \in \mathbb{R}^+$ *for the finite difference scheme and* $t^n = n\Delta t$ *for* $n \in \mathbb{N}_0$*. Let* $U^n_{j,k}$ *be the numerical solution at* (x_j, y_k, t^n)*, such that* $U^n_{j,k} \approx u(x_j, y_k, t^n)$ *for* $j = 0,\dots,N_x$ *and* $k = 0,\dots,N_y$ *and* $n \in \mathbb{N}$*. When* $n = 0$*,* U^0_j *is created by sampling* $u(x,y,0)$*, such that* $U^0_j := u(x_j, y_k, 0)$*, for* $j = 0,\dots,N_x-1$ *and* $k = 0,\dots,N_y-1$*. Define the vector* $\mathbf{U}^n \in \mathbb{R}^{N_xN_y}$ *where the* (j,k)*th element of* $\mathbf{U}^n$ *is defined by,*

$$\{\mathbf{U}^n\}_{(k-1)N_x+j} := U^n_{j-1,k-1}, \tag{5.6}$$

for $j = 1,\dots,N_x$ *and* $k = 1,\dots,N_y$*. This results in the following structure for* $\mathbf{U}^n$*,*

$$\mathbf{U}^n = \begin{bmatrix} U^n_{0,0} \\ \vdots \\ U^n_{N_x-1,0} \\ U^n_{0,1} \\ \vdots \\ U^n_{N_x-1,1} \\ \vdots \\ U^n_{0,N_y-1} \\ \vdots \\ U^n_{N_x-1,N_y-1} \end{bmatrix}. \tag{5.7}$$

We also define $h_1 := \frac{|\mu_1|\Delta t}{\Delta x}$ *and* $h_2 := \frac{|\mu_2|\Delta t}{\Delta y}$ *as the* CFL numbers in the x- and y-directions *respectively [13]. The* CFL number *for problem (5.1) is defined as* $h := h_1 + h_2$*. See Section 5.5.4 for details.*

In order to understand the structure of the vector $\mathbf{U}^n$, consider the vector $\mathbf{Q}^n_k \in \mathbb{R}^{N_x}$, such that its jth element is defined by $\{\mathbf{Q}^n_k\}_j = U^n_{j-1,k-1}$ for $j = 1,\dots,N_x$ and $k = 1,\dots,N_y$. The vector contains the numerical solution at each grid point in the x-direction, for fixed $y = y_k$, similar to the vector $\mathbf{U}^n$ in Section 3.3. These vectors are then stacked, to create the vector $\mathbf{U}^n$,

$$\mathbf{U}^n = [(\mathbf{Q}^n_0)^T, (\mathbf{Q}^n_1)^T, \dots, (\mathbf{Q}^n_{N_y-1})^T]^T. \tag{5.8}$$

The considered schemes are defined by the following schematics, when assuming $\mu_1, \mu_2 > 0$:

- The Upwind scheme (explicit scheme) [59, 13],

$$U_{j,k}^{n+1} = h_1 U_{j-1.k}^n + h_2 U_{j,k-1}^n + (1 - h_1 - h_2) U_{j,k}^n. \tag{5.9}$$

- The Crank-Nicolson scheme (implicit scheme) [70, 13],

$$\begin{aligned} &U_{j,k}^{n+1} + \frac{h_1}{4}\left(U_{j+1,k}^{n+1} - U_{j-1,k}^{n+1}\right) + \frac{h_2}{4}\left(U_{j,k+1}^{n+1} - U_{j,k-1}^{n+1}\right) \\ &= U_{j,k}^n - \frac{h_1}{4}\left(U_{j+1,k}^n - U_{j-1,k}^n\right) - \frac{h_2}{4}\left(U_{j,k+1}^n - U_{j,k-1}^n\right). \end{aligned} \tag{5.10}$$

We restrict our analysis to the case of $\mu_1, \mu_2 > 0$. These finite difference schemes can be applied to the discrete sample of the system found in the vector $\mathbf{U}^n$. This is achieved by constructing a matrix $M \in \mathbb{R}^{N_x N_y \times N_x N_y}$, using the schematics, that can be used to multiply $\mathbf{U}^n$ to create $\mathbf{U}^{n+1}$ as for the 1D linear advection problem. This advances the numerical solution at each grid point, forward Δt in time and results in $N = N_x N_y$, where N was defined in Section 2.3. Due to the circulant boundary conditions of problem (5.1), the matrix M implementing the above schemes is block circulant [65] and the blocks are circulant matrices.

In the case of the Upwind scheme, the schematic in (5.9) gives that if we set,

$$\begin{aligned} A &:= h_2 I_{N_x}, \;\; C := 0_{N_x}, \\ B &:= \begin{bmatrix} 1-h_1-h_2 & 0 & & \cdots & h_1 \\ h_1 & 1-h_1-h_2 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & h_1 & 1-h_1-h_2 & 0 \\ 0 & \cdots & 0 & h_1 & 1-h_1-h_2 \end{bmatrix}, \end{aligned} \tag{5.11}$$

where I_{N_x} and 0_{N_x} are the $N_x \times N_x$ identity and zero matrices respectively, then,

$$\begin{aligned} \mathbf{Q}_0^n &= A\mathbf{Q}_{N_y-1}^n + B\mathbf{Q}_0^n + C\mathbf{Q}_1^n, \\ \mathbf{Q}_k^n &= A\mathbf{Q}_{k-1}^n + B\mathbf{Q}_k^n + C\mathbf{Q}_{k+1}^n, \;\; \text{for } k = 1, \ldots, N_y - 2, \\ \mathbf{Q}_{N_y-1}^n &= A\mathbf{Q}_{N_y-2}^n + B\mathbf{Q}_{N_y-1}^n + C\mathbf{Q}_0^n, \end{aligned} \tag{5.12}$$

for all $n \in \mathbb{N}_0$. Using the structure of $\mathbf{U}^n$ in (5.8) this creates,

$$M = \begin{bmatrix} B & C & & \cdots & A \\ A & B & C & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & A & B & C \\ C & \cdots & 0 & A & B \end{bmatrix}, \tag{5.13}$$

so that $\mathbf{U}^{n+1} = M\mathbf{U}^n$ for all $n \in \mathbb{N}_0$. As $u(x_{N_x}, y, t^n) = u(x_0, y, t^n)$ for all $y \in [0, 1]$

and $n \in \mathbb{N}_0$, $U^n_{0,k-1} = U^n_{N_x,k-1}$ for all $k = 1, \ldots, N_y + 1$ and $n \in \mathbb{N}_0$. Similarly, as $u(x, y_{N_y}, t^n) = u(x, y_0, t^n)$ for all $x \in [0, 1]$ and $n \in \mathbb{N}_0$, $U^n_{j-1,0} = U^n_{j-1,N_y}$ for all $j = 1, \ldots, N_x + 1$ and $n \in \mathbb{N}_0$.

5.3.1 The 2D discrete Fourier transform

As the vector $\mathbf{U}^n$ is an N_xN_y-dimensional, it can be constructed from the N_xN_y vectors of the *2D DFT* basis [60] $\{\mathbf{v}_{p,q}\}_{p=1,q=1}^{N_x,N_y}$, such that,

$$\begin{aligned} \{\mathbf{v}_{p,q}\}_{(k-1)N_x+j} &= \frac{1}{\sqrt{N_xN_y}} e^{\frac{2\pi i(p-1)(j-1)}{N_x}} e^{\frac{2\pi i(q-1)(k-1)}{N_y}}, \\ &= \frac{1}{\sqrt{N_xN_y}} e^{2\pi i(p-1)x_{j-1}} e^{2\pi i(q-1)y_{k-1}}, \end{aligned} \tag{5.14}$$

for $p, j = 1, \ldots, N_x$ and $k, q = 1, \ldots, N_y$. This is the $(p-1, q-1)$th 2D Fourier basis function sampled at (x_{j-1}, y_{k-1}) in space, with amplitude $\frac{1}{\sqrt{N_xN_y}}$. The vectors form an orthonormal basis for $\mathbb{R}^{N_xN_y}$ [60], ie: $\mathbf{v}^*_{p,q}\mathbf{v}_{r,s} = \delta_{p,q}\delta_{r,s}$. The numerical solution is constructed from N_xN_y 2D Fourier basis functions. These vectors form an orthonormal set of eigenvectors for the matrix M, for the two considered schemes. The eigenvalue decomposition for the matrix M for each scheme can be constructed using thesis eigenvectors,

$$M = V\Lambda V^{-1} = V\Lambda V^*. \tag{5.15}$$

As before, the matrix $V \in \mathbb{C}^{N_xN_y \times N_xN_y}$ is constructed from the 2D DFT eigenbasis, such that the $\{(q-1)N_x+p\}$th column of V is $\mathbf{v}_{p,q}$ for $p = 1, \ldots, N_x$ and $q = 1, \ldots, N_y$. Then the elements of V are defined as,

$$\{V\}_{(k-1)N_x+j,(q-1)N_x+p} = \frac{1}{\sqrt{N_xN_y}} e^{\frac{2\pi i(p-1)(j-1)}{N_x}} e^{\frac{2\pi i(q-1)(k-1)}{N_y}}, \tag{5.16}$$

for $j, p = 1, \ldots, N_x$ and $k, q = 1, \ldots, N_y$. It is a unitary matrix. The matrix $\Lambda \in \mathbb{C}^{N_x,N_y \times N_xN_y}$ is the diagonal matrix of eigenvalues corresponding to the eigenvectors in the matrix V, for the chosen scheme. The eigenvalue $\lambda_{p,q}$, corresponding to $\mathbf{v}_{p,q}$ is found in $\{\Lambda\}_{(q-1)N_x+p,(q-1)N_x+p}$ for $p = 1, \ldots, N_x$ and $q = 1, \ldots, N_y$. This means that $\lambda_{p,q}$ corresponds to the $(p-1, q-1)$th wavenumber component of the numerical solution. As for the 1D problem, the eigenvalues of the scheme are scheme dependent whilst the eigenvectors are scheme independent.

As for the 1D case, using the 2D DFT basis to construct the state of the numerical solution results in the construction of a *2D discrete Fourier series*,

$$\{\mathbf{U}^0\}_{(k-1)N_x+j} = \sum_{p=1}^{N_x}\sum_{q=1}^{N_y} \alpha_{p,q} e^{\frac{2\pi i(p-1)(j-1)}{N_x}} e^{\frac{2\pi i(q-1)(k-1)}{N_y}}, \tag{5.17}$$

for $j = 1, \ldots, N_x$ and $k = 1, \ldots, N_y$. Performing the same analysis as in Section 3.3.1,

we find that by applying $\mathbf{v}_{p,q}^*$ to $\mathbf{U}^0$, we obtain the coefficient $\alpha_{p,q}$ for $p = 1, \ldots, N_x$ and $q = 1, \ldots, N_y$. Therefore $\mathbf{v}_{p,q}^*$ applies the 2D DFT for the $(p-1, q-1)$th wavenumber component of the 2D discrete Fourier series. As $\mathbf{v}_{p,q}$ forms the columns of V, applying V^* to any vector $\mathbf{z} \in \mathbb{R}^{N_x N_y}$ applies the 2D DFT, identifying the coefficients for the 2D discrete Fourier series constructing $\mathbf{z}$. Applying V to $V^*\mathbf{z}$, reconstructs $\mathbf{z}$ using a 2D discrete Fourier series as $VV^*\mathbf{z} = \mathbf{z}$. Therefore, V applies the 2D IDFT.

Define the operator $\mathcal{F} : \mathbb{R}^{N_x N_y} \to \mathbb{C}^{N_x N_y}$, $\mathbf{z} \mapsto \mathcal{F}(\mathbf{z}) = V^*\mathbf{z}$, to implement the 2D DFT. Denote the $\{(q-1)N_x+p\}$th element of $\mathcal{F}(\mathbf{z})$, the coefficient for the $(p-1, q-1)$th wavenumber component, by $\mathcal{F}_{p,q}(\mathbf{z}) = \{V^*\mathbf{z}\}_{(q-1)N_x+p} = \mathbf{v}_{p,q}^*\mathbf{z}$ for $p = 1, \ldots, N_x$ and $q = 1, \ldots, N_y$. Using this definition in (5.17) results in $\mathcal{F}_{p,q}(\mathbf{z}) = \alpha_{p,q}$ for $p = 1, \ldots, N_x$ and $q = 1, \ldots, N_y$.

Consider the nth state of the numerical model given by $\mathbf{U}^n = M^n\mathbf{U}^0$. Applying the 2D DFT to $\mathbf{U}^n$,

$$\begin{aligned} & \mathcal{F}(\mathbf{U}^n) = V^*V\Lambda^n V^*\mathbf{U}^0 = \Lambda^n \mathcal{F}(\mathbf{U}^0) \\ \Rightarrow \quad & \mathcal{F}_{p,q}(\mathbf{U}^n) = \lambda_{p,q}^n \mathcal{F}_{p,q}(\mathbf{U}^0) = \lambda_{p,q}^n \alpha_{p,q}, \end{aligned} \tag{5.18}$$

for $p = 1, \ldots, N_x$ and $q = 1, \ldots, N_y$. This gives that the coefficient for the eigenvector $\mathbf{v}_{p,q}$ in $\mathbf{U}^n$ is $\lambda_{p,q}^n \alpha_{p,q}$, for $p = 1, \ldots, N_x$ and $q = 1, \ldots, N_y$. Just as with the 1D problem, it is the eigenvalues of the matrix M which propagate the state of the system forward Δt through time. Then any errors introduced into the numerical solution, are due to errors in these eigenvalues.

As with the 1D DFT basis, the complex conjugate of each eigenvector $\mathbf{v}_{p,q}$ in the 2D DFT basis, is also an vector in the basis,

- $\mathcal{F}_{1,1}(\mathbf{z}) \in \mathbb{R}$,
- $\overline{\mathcal{F}_{1,q}(\mathbf{z})} = \mathcal{F}_{1,N_y-q+2}(\mathbf{z})$, for $q = 2, \ldots, N_y$,
- $\overline{\mathcal{F}_{p,1}(\mathbf{z})} = \mathcal{F}_{N_x-p+2,1}(\mathbf{z})$, for $p = 2, \ldots, N_x$,
- $\overline{\mathcal{F}_{p,q}(\mathbf{z})} = \mathcal{F}_{N_x-p+2,N_y-q+2}(\mathbf{z})$, for $p = 2, \ldots, N_x$ and $q = 2, \ldots, N_y$,

for all $\mathbf{z} \in \mathbb{R}^{N_x N_y}$. When N_x and N_y are both even, $\mathbf{v}_{\frac{N_x}{2}+1, \frac{N_y}{2}+1}$ is also real, hence $\mathcal{F}_{\frac{N_x}{2}+1, \frac{N_y}{2}+1}(\mathbf{z})$ is also real for all $\mathbf{z} \in \mathbb{R}^{N_x N_y}$.

The complex conjugate nature of the coefficients found through the 2D DFT, means that by examining (5.18), we find that $\lambda_{p,q}$ has the same complex conjugate pairing. When N_x and N_y are both even, $\lambda_{\frac{N_x}{2}+1, \frac{N_y}{2}+1} \in \mathbb{R}$. Since the eigenvalues are complex, we write them in polar co-ordinate form as $\lambda_{p,q} = |\lambda_{p,q}|e^{i\theta_{p,q}}$, where $\theta_{p,q} \in [-\pi, \pi)$, such that,

- $\theta_{1,1} = 0$,
- $-\theta_{1,q} = \theta_{1,N_y-q+2}$, for $q = 2, \ldots, N_y$,

- $-\theta_{p,1} = \theta_{N_x-p+2,1}$, for $p = 2, \ldots, N_x$,
- $-\theta_{p,q} = \theta_{N_x-p+2,N_y-q+2}$, for $p = 2, \ldots, N_x$ and $q = 2, \ldots, N_y$.

Summing the complex conjugate wavenumber components of a discrete Fourier series creates a *2D real wavenumber component*. The state of the system is then constructed from $\left\lfloor \frac{N_x N_y}{2} \right\rfloor + 1$ 2D real wavenumber components.

The eigenvector $\mathbf{v}_{p,q}$ has been discussed as corresponding to the $(p-1, q-1)$th 2D Fourier basis function, sampled at the grid points of the finite difference scheme. However as with the 1D DFT basis, the wavenumber components for $p = \left\lfloor \frac{N_x}{2} \right\rfloor + 2, \ldots, N_x$ play the role of the negative wavenumber components of the 2D Fourier series in the x-wavenumber. Similarly, the wavenumber components with $q = \left\lfloor \frac{N_y}{2} \right\rfloor + 2, \ldots, N_y$ play the role of the negative wavenumber components of the 2D Fourier series in the y-wavenumber. This is due to the same aliasing effects in each direction as discussed in Section 3.4. Therefore when $p = \left\lfloor \frac{N_x}{2} \right\rfloor + 2, \ldots, N_x$ the wavenumber components corresponds to the $(-N_x + p - 1)$th wavenumber component in the x-direction, of a 2D Fourier series. Similarly, when $q = \left\lfloor \frac{N_y}{2} \right\rfloor + 2, \ldots, N_y$, the wavenumber component corresponds to the $(-N_y + q - 1)$th wavenumber component in the y-direction, of a 2D Fourier series.

5.3.2 Aliasing and the Poisson summation in 2D

Aliasing occurs in the 2D problem, just as for the 1D problem. The Fourier basis functions of a 2D Fourier series are composed of the Fourier basis functions for a 1D Fourier series,

$$e^{2\pi i p x} \times e^{2\pi i q y}, \tag{5.19}$$

for $p, q \in \mathbb{Z}$. This means that aliasing of the 2D Fourier basis functions in 2D Fourier series can be discussed in terms of aliasing in the x- and y-directions individually,

$$e^{2\pi i p x_{j-1}} e^{2\pi i q x_{k-1}} = e^{\frac{2\pi i p(j-1)}{N_x}} e^{\frac{2\pi i q(k-1)}{N_y}} = e^{2\pi i [p]_{N_x} x_{j-1}} e^{2\pi i [q]_{N_y} y_{k-1}}, \tag{5.20}$$

for $p, q \in \mathbb{Z}$, $j = 1, \ldots, N_x$ and $k = 1, \ldots, N_y$. Here $[\cdot]$ denotes modulo with respect to the given subscript. As a result the analysis of Section 3.4 can be used to determine the Nyquist rate of the x- and y components of the 2D wavenumber components, independently.

The Poisson summation in 2D is found similarly to the 1D case. Let $u_0(x, y)$ be continuous at each sample point on the grid and possess a convergent 2D Fourier series.

Then we can represent $u_0(x, y)$ at each sample point by its 2D Fourier series and obtain,

$$\begin{aligned}
& \mathcal{F}_{p,q}(\mathbf{U}^0), \\
= \; & \mathbf{v}^*_{(q-1)N_x+p}\mathbf{U}^0, \\
= \; & \frac{1}{\sqrt{N_xN_y}} \sum_{s=1}^{N_x}\sum_{r=1}^{N_y} u(x_{s-1}, y_{r-1}, t^0) e^{\frac{-2\pi i(p-1)(s-1)}{N_x}} e^{\frac{-2\pi i(q-1)(r-1)}{N_y}}, \\
= \; & \frac{1}{\sqrt{N_xN_y}} \sum_{s=1}^{N_x}\sum_{r=1}^{N_y}\sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} c_{j,k} e^{\frac{2\pi ij(s-1)}{N_x}} e^{\frac{2\pi ik(r-1)}{N_y}} e^{\frac{-2\pi i(p-1)(s-1)}{N_x}} e^{\frac{-2\pi i(q-1)(r-1)}{N_y}}, \\
= \; & \sqrt{N_xN_y} \sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} c_{p-1+jN_x, q-1+kN_y}. \qquad (5.21)
\end{aligned}$$

This shows that the coefficient of the $(p-1, q-1)$th resolvable wavenumber component found through the 2D DFT, is made up of the coefficients of the wavenumber components from the 2D Fourier series of $u_0(x, y)$, which are aliased to the $(p-1, q-1)$th wavenumber component. When $u_0(x, y)$ has a discontinuous sample point, we represent it using the Fourier series of an alternative function, which is continuous and equal to the original function at each sample point and has a convergent Fourier series.

Applying a finite difference scheme implemented by the matrix M, to the vector $\mathbf{U}^0$ to progress the numerical solution Δt in time results in,

$$\mathcal{F}_{p,q}(\mathbf{U}^1) = \lambda_{p,q}\mathcal{F}_{p,q}(\mathbf{U}^0) = \sqrt{N_xN_y} \sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} \lambda_{p,q} c_{p-1+jN_x, q-1+kN_y}. \qquad (5.22)$$

So we get a similar result to the 1D Poisson summation. The eigenvalue $\lambda_{p,q}$ of the scheme, propagates the 2D wavenumber components of the solution with wavenumbers $(p-1+jN_x, q-1+kN_y)$, for $p = 1, \ldots, N_x$, $q = 1, \ldots, N_y$ and $j, k \in \mathbb{Z}$. This allows M to propagate all wavenumber components of the numerical solution by only directly acting on N_xN_y of them.

Then as before, M applies the same magnitude and phase shifts to an unresolvable wavenumber component, as it does to the resolvable wavenumber component they alias to. If the magnitude and phase shift for the resolvable wavenumber component is correct, this does not necessarily mean that this is the correct magnitude and phase shift for the unresolvable 2D wavenumber component.

5.4 Numerical dissipation and dispersion in 2D

The similarities between the implementation of the numerical solutions for problems (5.1) and (3.1), means that the numerical model error introduced into the numerical solution of the 2D problem, can be examined in the same way as was done for the 1D problem. Hence we can consider the error introduced into the solution in terms of

numerical dissipation and dispersion, when the considered finite difference schemes are numerically stable.

Numerical dissipation and dispersion are introduced by the eigenvalues of the scheme as before. The definitions of numerical dissipation and dispersion in Definitions 3.4 and 3.5 respectively in Section 3.5 hold for the 2D problem, but with the examples modified to reflect that the eigenvalues for the 2D problem have two indices. Using these definitions requires comparing the coefficients of the Fourier series for the analytical solution with the coefficients of the 2D Fourier series for the numerical solution. This is done by investigating how the magnitude and phase of these coefficients change in time Δt. To this end we define the Fourier series for the numerical solution using (5.22). Define the function $w : \mathbb{R} \times \mathbb{R} \times [0, \infty) \to \mathbb{R}$ such that,

$$(x, y, t) \mapsto w(x, y, t) = \sum_{p=-\infty}^{\infty} \sum_{q=-\infty}^{\infty} v_{p,q}(t) e^{2\pi i p x} e^{2\pi i q y}, \tag{5.23}$$

where $v_{p,q} : [0, \infty) \to \mathbb{C}$ such that,

$$t \mapsto v_{p,q}(t) = \lambda_{[p]_{N_x}+1,[q]_{N_y}+1}^{\frac{t}{\Delta t}} v_{p,q}(0), \quad v_{p,q}(0) = c_{p,q}, \quad \forall p, q \in \mathbb{Z}. \tag{5.24}$$

Using this definition, $w(x, y, 0)$ is equal to the Fourier series of $u(x, y, 0)$ for all $x, y \in \mathbb{R}$. Evaluating $w(x_j, y_k, t^n)$ for some $j = 0, \ldots, N_x - 1$, $k = 0, \ldots, N_y - 1$ and $n \in \mathbb{N}_0$, produces the state of the numerical solution at these points. At non-integer multiples of Δx, Δy or Δt, the numerical solution is interpolated in the corresponding variable.

Define the function $g_{p,q}^{scheme} : \mathbb{C} \to \mathbb{C}$ that maps the coefficients of the Fourier series in (5.23), Δt through time,

$$z \mapsto g_{p,q}^{scheme}(z) = \lambda_{[p]_{N_x}+1,[q]_{N_y}+1} z. \tag{5.25}$$

The function that maps the coefficients of the Fourier series in (5.4) Δt through time, is defined by $g_{p,q} : \mathbb{C} \to \mathbb{C}$, such that $b_{p,q}(n\Delta t) \mapsto g_{p,q}(b_{p,q}(n\Delta t)) = b_{p,q}((n+1)\Delta t)$ for all $n \in \mathbb{N}_0$. Suppose the functions $b_{p,q}(t)$ are invertible for all (p, q), then $g_{p,q}(\cdot)$ is defined as,

$$g_{p,q}(\cdot) := b_{p,q}(b_{p,q}^{-1}(\cdot) + \Delta t). \tag{5.26}$$

As for the 1D scheme, it can be shown that $g_{p,q}^{scheme}(\cdot)$ can also be defined in this form, using $v_{p,q}(\cdot)$ instead of $b_{p,q}(\cdot)$. The definitions of numerical dissipation and dispersion in Definition 3.4 and 3.5 respectively require that the magnitude and phase of $g_{p,q}^{scheme}(1) = \lambda_{[p]_{N_x}+1,[q]_{N_y}+1}$ be compared with the same for $g_{p,q}(1)$, to determine the numerically dissipative and dispersive properties of the scheme, respectively.

5.4.1 The Fourier series solution to the 2D linear advection problem

The Fourier series solution to problem (5.1) is found through substituting the solution $e^{2\pi i(px+qy-\omega t)}$ into the problem. Given the coefficients $c_{p,q}$ for the Fourier series of $u_0(x,y)$ in (5.5), the Fourier series for $u(x,y,t)$ is given by,

$$u(x,y,t) \sim \sum_{p=-\infty}^{\infty} \sum_{q=-\infty}^{\infty} c_{p,q} e^{2\pi ip(x-\mu_1 t)} e^{2\pi iq(y-\mu_2 t)}. \tag{5.27}$$

Comparing this with the Fourier series for $u(x,y,t)$ in (5.4) finds that,

$$b_{p,q}(t) = c_{p,q} e^{-2\pi ip\mu_1 t} e^{-2\pi iq\mu_2 t} = b_{p,q}(0) e^{-2\pi ip\mu_1 t} e^{-2\pi iq\mu_2 t},$$

and the time dependent portion of this coefficient is $e^{-2\pi ip\mu_1 t} e^{-2\pi iq\mu_2 t}$. An interesting property of this coefficient is that it can be decomposed into a function of x multiplied by a function of y. The function $g_{p,q}(\cdot)$ defined in Section 5.4, is then given by,

$$g_{p,q}(z) = e^{-2\pi ip\mu_1 \Delta t} e^{-2\pi iq\mu_2 \Delta t} z. \tag{5.28}$$

Applying $g_{p,q}(\cdot)$ to $b_{p,q}(t)$ moves the coefficient Δt through time to $b_{p,q}(t+\Delta t)$.

Performing a similar analysis to Remark 3.6 in Section 3.5.1,

$$g_{p,q}(1) = e^{-2\pi ip\mu_1 \Delta t} e^{-2\pi iq\mu_2 \Delta t},$$

is analysed for problem (5.1) to reveal that a finite difference scheme for solving problem (5.1) is:

- numerically non-dissipative and non-dispersive with respect to the resolvable wavenumber components, aliasing will still occur if $h_1, h_2 \in \mathbb{R}^+ \setminus \mathbb{N}$, but this will be a form of numerical dispersion (MNIMC scheme - see Section 5.6.1),
- numerically dissipative and non-dispersive with respect to the resolvable wavenumber components, aliasing will be a form of numerical dissipation, but will also be a form of numerical dispersion if $h \in \mathbb{R}^+ \setminus \mathbb{N}$ (no scheme is considered with this property for solving problem (5.1) in this thesis),
- numerically non-dissipative and dispersive with respect to the resolvable wavenumber components, aliasing will be a form of numerical dispersion (Crank-Nicolson scheme for $0 < h < 1$),
- numerically dissipative and dispersive with respect to the resolvable wavenumber components, aliasing will be a form of numerical dissipation and dispersion (Upwind scheme for $0 < h < 1$).

5.5 Analysis of finite difference schemes for the 2D linear advection problem

As for the 1D linear advection problem, before using either the Upwind or Crank-Nicolson schemes to solve the 2D linear advection problem, we need to identify when the schemes are convergent. This relies on checking the consistency and numerical stability of each scheme in Section 5.3. These properties are presented in Table 5.1. The numerically dissipative and dispersive properties also need to be identified and can be seen in Table 5.2, for when the schemes are numerically stable. These properties can be identified through the eigenvalues of the schemes.

5.5.1 The 2D Upwind scheme

The Upwind scheme in (5.9) is an explicit finite difference scheme, derived to solve the 2D linear advection problem, by approximating the temporal derivative using a forward difference in time and the spatial derivatives by backward differences in space. This scheme is only numerically stable for $\mu_1, \mu_2 > 0$, similarly to the reasons discussed in Section 3.6.1.

The eigenvalues of the matrix implementing the Upwind scheme are,

$$\begin{aligned}\lambda_{p,q} &= 1 + h_1\left\{\cos\left[\frac{2\pi(p-1)}{N_x}\right] - 1\right\} + h_2\left\{\cos\left[\frac{2\pi(q-1)}{N_y}\right] - 1\right\} \\ &\quad -i\left\{h_1 \sin\left[\frac{2\pi(p-1)}{N_x}\right] + h_2 \sin\left[\frac{2\pi(q-1)}{N_y}\right]\right\}, \end{aligned} \tag{5.29}$$

for $p = 1, \ldots, N_x$ and $q = 1, \ldots, N_y$.

5.5.2 The Crank-Nicolson scheme

The Crank-Nicolson scheme in (5.10) is an implicit scheme derived to solve the 2D linear advection problem, by using a forward difference in time to approximate the temporal partial derivative and the average of two central differences in space, to approximate the partial derivatives in each space dimension. In the case of the partial derivative with respect to x, a central difference at x_j is calculated at times t^{n+1} and t^n. The partial derivative with respect to x is then calculated from the average of these central differences. The same method is used to approximate the partial derivative with respect to y at y_k [70, 13].

The eigenvalues of the matrix implementing the Crank-Nicolson scheme are,

$$\lambda_{p,q} = \frac{1 - \left\{\frac{h_1}{2}\sin\left[\frac{2\pi(p-1)}{N_x}\right] + \frac{h_2}{2}\sin\left[\frac{2\pi(q-1)}{N_y}\right]\right\}^2 - i\left\{h_1 \sin\left[\frac{2\pi(p-1)}{N_x}\right] + h_2 \sin\left[\frac{2\pi(q-1)}{N_y}\right]\right\}}{1 + \left\{\frac{h_1}{2}\sin\left[\frac{2\pi(p-1)}{N_x}\right] + \frac{h_2}{2}\sin\left[\frac{2\pi(q-1)}{N_y}\right]\right\}^2}, \tag{5.30}$$

for $p = 1, \ldots, N_x$ and $q = 1, \ldots, N_y$. We note that $|\lambda_{p,q}| = 1$ for $p = 1, \ldots, N_x$ and $q = 1, \ldots, N_y$, so the scheme is always numerically stable and numerically non-dissipative.

5.5.3 Two-dimensional finite difference scheme property summary

Tables 5.1 and 5.2 summarise the properties of the Upwind and Crank-Nicolson schemes as identified through the eigenvalues of the schemes in Sections 5.5.1 and 5.5.2. The properties of the NIMC and MNIMC schemes for the 2D linear advection problem defined in Sections 5.6 and 5.6.1 respectively, are also included in the tables for comparison. The numerically dissipative and dispersive properties of the schemes were determined similarly to that of the schemes for the 1D linear advection problem, by defining the similar continuous variables in the x- and y-directions, then using differentiation.

Scheme	Consistent	Numerically Stable	Convergent	Singular Matrix
Upwind	Always	$0 < h \leq 1$	$0 < h \leq 1$	N_x is even and $h_1 = h_2 = \frac{1}{2}$
Crank-Nicolson	Always	Always	Always	Never
NIMC	$h = 1$	$h = 1$	$h = 1$	Never
MNIMC	Always	Always	Always	Never

Table 5.1: *This Table summarises the consistency, numerical stability and hence convergence properties, for the finite difference schemes considered for solving problem (5.1). The consistency of the scheme is for sufficiently smooth initial conditions. Information on the invertibility of the matrix used to implement the scheme is also provided.*

Scheme	Non-Dissipative wrt resolvable wavenumber components	Non-Dispersive wrt resolvable wavenumber components	Non-Dissipative wrt all wavenumber components	Non-Dispersive wrt all wavenumber components
Upwind	Never	Never	Never	Never
Crank-Nicolson	Always	Never	Always	Never
NIMC	$h = 1$	$h = 1$	$h = 1$	$h = 1$
MNIMC	Always	Always	Always	$h_1, h_2 \in \mathbb{N}$

Table 5.2: *This Table summarises the numerically dissipative and dispersive properties with respect to the resolvable wavenumber components and all wavenumber components of the numerical solution, for the finite difference schemes considered for solving problem (5.1), for* $0 < h \leq 1$. *Here 'wrt' stands for 'with respect to'.*

5.5.4 The CFL condition

The CFL condition for problem (5.1) is derived as discussed in Section 3.6.5. However, it is difficult to extract the form of the CFL for a two-dimensional problem in space, from the paper by Courant et al. [73]. The CFL number for a two-dimensional problem is stated for specific problems in literature such as [13], without derivation. In order to verify our understanding of the CFL number for a two-dimensional problem, we present a derivation of this quantity for completeness. This is not a new conclusion on the CFL number. The CFL condition for a two-dimensional problem is derived similarly to that of the CFL condition in Section 3.6.5, for a one-dimensional problem. We require that for a explicit finite difference scheme to have the possibility of converging to the solution of the PDE as $\Delta t, \Delta x, \Delta y \to 0$, the domain of dependence of the PDE must lie within the domain of dependence of the numerical scheme [73, 14]. This is the same requirement as for the 1D case and leads to the CFL condition for two-dimensional problems also being a necessary condition for the convergence of explicit finite difference schemes in two-dimensions.

Examining the Upwind and Crank-Nicolson schemes in (5.9) and (5.10) respectively, we can see that the quantities $h_1 = \frac{\mu_1 \Delta t}{\Delta x}$ and $h_2 = \frac{\mu_2 \Delta t}{\Delta y}$ play a role in the schemes ($\mu_1, \mu_2 > 0$). Comparing them with the CFL number for the 1D linear advection problem, they appear to be the CFL numbers for this problem in the x- and y-directions respectively. However, this does not help us identify the CFL condition and hence the CFL number for a two-dimensional problem, but does tell us that they are likely to play a role in it.

In order to demonstrate the derivation of the CFL condition for a two-dimensional problem, consider the Upwind scheme in (5.9), for solving problem (5.1). Consider the point (x_j, y_k, t^{n+1}). The Upwind scheme calculates a numerical approximation to the solution of problem (5.1) at $u(x_j, y_k, t^{n+1})$, given by $U_{j,k}^{n+1}$, using the points $U_{j-1,k}^{n}$, $U_{j,k}^{n}$ and $U_{j,k-1}^{n}$. Following all the data points used to create $U_{j,k}^{n+1}$, backwards in time to time $t = 0$, creates the domain of dependence for the Upwind scheme. This is a pyramidal structure as shown in Figure (5.1). The plane forming the sloping face of the pyramid, is given by,

$$\begin{bmatrix} x \\ y \\ t \end{bmatrix} = \begin{bmatrix} x_j \\ y_k \\ t^{n+1} \end{bmatrix} + s_1 \begin{bmatrix} -\Delta x \\ 0 \\ -\Delta t \end{bmatrix} + s_2 \begin{bmatrix} 0 \\ -\Delta y \\ -\Delta t \end{bmatrix}, \tag{5.31}$$

for $s_1, s_2 \in \mathbb{R}^+$.

The domain of dependence for the PDE through the point (x_j, y_k, t^{n+1}), is the characteristic line for the PDE that passes through this point,

$$x = \mu_1(t - t^{n+1}) + x_j \text{ and } y = \mu_2(t - t^{n+1}) + y_k, \tag{5.32}$$

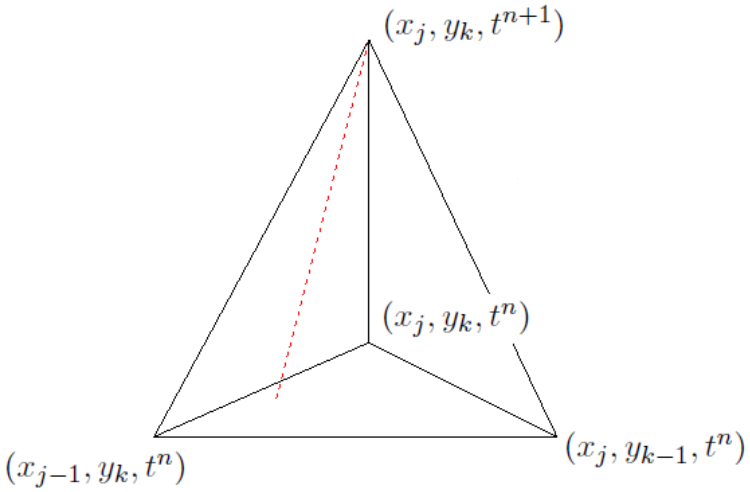


Figure 5.1: *The prism with a solid boundary forms the domain of dependence for the Upwind finite difference scheme in (5.9). The red dotted line forms the domain of dependence for the 2D linear advection problem in (5.1), through the point (x_j, y_k, t^{n+1}). The 2D CFL condition requires that the domain of dependence of the PDE be contained within the domain of dependence of the finite difference scheme.*

for $t \in [0, \infty)$. This gives that given any time t, the characteristics in the x- and y-directions are independent. Then the equation of the characteristic line is,

$$\begin{bmatrix} x \\ y \\ t \end{bmatrix} = \begin{bmatrix} x_j \\ y_k \\ t^{n+1} \end{bmatrix} + c \begin{bmatrix} -\mu_1 \Delta t \\ -\mu_2 \Delta t \\ -\Delta t \end{bmatrix}, \tag{5.33}$$

for $c \in [0, \infty)$.

We require that the characteristic line lies within the domain of dependence of the numerical scheme. This means that for any time t, the x and y points of the characteristic curve must line within the domain of dependence of the numerical scheme. Then by (5.31) and (5.33),

$$x_j - s_1 \Delta x \leq x_j - c\mu_1 \Delta t \quad \Rightarrow \quad s_1 \geq c h_1 \tag{5.34}$$

$$y_k - s_2 \Delta y \leq y_k - c\mu_2 \Delta t \quad \Rightarrow \quad s_2 \geq c h_2 \tag{5.35}$$

$$t^{n+1} - (s_1 + s_2)\Delta t = t^{n+1} - c\Delta t \quad \Rightarrow \quad c = s_1 + s_2 \tag{5.36}$$

Combining these equations reveals that we require that,

$$c \geq c h_1 + c h_2 \Rightarrow 1 \geq h_1 + h_2 \tag{5.37}$$

as $c \geq 0$. This provides the CFL condition for the Upwind scheme for solving problem (5.1). The Upwind scheme is only valid when $\mu_1, \mu_2 > 0$.

5.6 Generating perfect observations for the 2D linear advection problem

In this Section we wish to investigate the creation of perfect observations for use in our strong constraint 4D-Var data assimilation problem, where our physical system is defined as the 2D linear advection problem. This problem is an extension of the 1D linear advection problem, so we can extend many of the ideas developed in Chapter 3, to generate perfect observations for the 2D problem.

In the 1D linear advection problem, we were able to make use of MATLABs®[74] *circshift* function to numerically generate observations, due to the form of the analytical solution. The analytical solution for the 2D problem has similar properties, preserving the shape of the initial condition over time, as it is propagated with constant wave speeds in both the x- and y-directions. As a result, we can again make use of the *circshift* function to numerically generate perfect observations.

Analytical observations for the 1D linear advection problem were defined in terms of the MNIMC scheme for the 1D linear advection problem, a numerically non-dissipative and non-dispersive finite difference scheme with respect to the resolvable wavenumber components, plus an additive correction term for the aliasing errors introduced by the MNIMC scheme. We wish to do the same for the 2D linear advection problem. The MNIMC scheme for the 1D linear advection problem was initially derived from the NIMC scheme for the problem. The NIMC scheme for the 2D linear advection problem is defined by,

$$U_{j,k}^{n+1} = hU_{j-1,k-1}^{n}, \tag{5.38}$$

and is implemented via the matrix $M_{NIMC} \in \mathbb{R}^{N_x N_y \times N_x N_y}$ such that $\mathbf{U}^{n+1} = M_{NIMC}\mathbf{U}^{n}$. The 2D DFT basis forms the eigenvectors for the matrix M_{NIMC}. This scheme possesses similar limitations to that of the NIMC scheme for the 1D linear advection problem, requiring $h_{NIMC} = 1$ in order to converge to a solution for the 2D linear advection problem, as can be seen in Table 5.1. Instead of deriving the MNIMC scheme via the NIMC scheme, we go straight to developing the MNIMC scheme for the 2D linear advection problem using the Fourier series method developed in Section 3.7.5, due to the simplicity of this method.

5.6.1 The MNIMC scheme for the 2D linear advection problem

The MNIMC scheme for the 2D linear advection problem is derived using the ideas set out in Section 3.7.5. Let the vectors of the 2D DFT basis defined in Section 5.3, form the eigenvectors for the scheme. Define the eigenvalues of the scheme by $\left\{\tilde{\lambda}_{p,q}\right\}_{p=1,q=1}^{N_x,N_y}$ where $\tilde{\lambda}_{p,q} \in \mathbb{C}$ is the eigenvalues corresponding the eigenvectors $\mathbf{v}_{p,q}$. The eigenvalue $\tilde{\lambda}_{p,q}$ is defined so that it correctly propagates the resolvable wavenumber component of the Fourier series for the numerical solution in (5.23), that $\mathbf{v}_{p,q}$ corresponds to.

Therefore we define $\tilde{\lambda}_{p,q}$ using the corresponding $g_{j,k}(1) = e^{-2\pi i j \mu_1 \Delta t} e^{-2\pi i k \mu_2 \Delta t}$. This results in,

$$\tilde{\lambda}_{p,q} = \begin{cases} g_{p-1,q-1}(1), & \text{for } p = 1,\ldots,\left\lfloor \frac{N_x}{2} \right\rfloor + 1 \\ & \text{and } q = 1,\ldots,\left\lfloor \frac{N_y}{2} \right\rfloor + 1, \\ g_{p-1,-N_y+q-1}(1), & \text{for } p = 1,\ldots,\left\lfloor \frac{N_x}{2} \right\rfloor + 1 \\ & \text{and } q = \left\lfloor \frac{N_y}{2} \right\rfloor + 2,\ldots,N_y, \\ g_{-N_x+p-1,q-1}(1), & \text{for } p = \left\lfloor \frac{N_x}{2} \right\rfloor + 2,\ldots,N_x \\ & \text{and } q = 1,\ldots,\left\lfloor \frac{N_y}{2} \right\rfloor + 1, \\ g_{-N_x+p-1,-N_y+q-1}(1), & \text{for } p = \left\lfloor \frac{N_x}{2} \right\rfloor + 2,\ldots,N_x \\ & \text{and } q = \left\lfloor \frac{N_y}{2} \right\rfloor + 2,\ldots,N_y, \end{cases} \tag{5.39}$$

$$= \begin{cases} e^{-2\pi i(p-1)\mu_1 \Delta t} e^{-2\pi i(q-1)\mu_2 \Delta t}, & \text{for } p = 1,\ldots,\left\lfloor \frac{N_x}{2} \right\rfloor + 1 \\ & \text{and } q = 1,\ldots,\left\lfloor \frac{N_y}{2} \right\rfloor + 1, \\ e^{-2\pi i(p-1)\mu_1 \Delta t} e^{-2\pi i(-N_y+q-1)\mu_2 \Delta t}, & \text{for } p = 1,\ldots,\left\lfloor \frac{N_x}{2} \right\rfloor + 1 \\ & \text{and } q = \left\lfloor \frac{N_y}{2} \right\rfloor + 2,\ldots,N_y, \\ e^{-2\pi i(-N_x+p-1)\mu_1 \Delta t} e^{-2\pi i(q-1)\mu_2 \Delta t}, & \text{for } p = \left\lfloor \frac{N_x}{2} \right\rfloor + 2,\ldots,N_x \\ & \text{and } q = 1,\ldots,\left\lfloor \frac{N_y}{2} \right\rfloor + 1, \\ e^{-2\pi i(-N_x+p-1)\mu_1 \Delta t} e^{-2\pi i(-N_y+q-1)\mu_2 \Delta t}, & \text{for } p = \left\lfloor \frac{N_x}{2} \right\rfloor + 2,\ldots,N_x \\ & \text{and } q = \left\lfloor \frac{N_y}{2} \right\rfloor + 2,\ldots,N_y, \end{cases}$$

$$= \begin{cases} e^{\frac{-2\pi i(p-1)\mathrm{sgn}(\mu_1)h_1}{N_x}} e^{\frac{-2\pi i(q-1)\mathrm{sgn}(\mu_2)h_2}{N_y}}, & \text{for } p = 1,\ldots,\left\lfloor \frac{N_x}{2} \right\rfloor + 1 \\ & \text{and } q = 1,\ldots,\left\lfloor \frac{N_y}{2} \right\rfloor + 1, \\ e^{\frac{-2\pi i(p-1)\mathrm{sgn}(\mu_1)h_1}{N_x}} e^{\frac{2\pi i(N_y-q+1)\mathrm{sgn}(\mu_2)h_2}{N_y}}, & \text{for } p = 1,\ldots,\left\lfloor \frac{N_x}{2} \right\rfloor + 1 \\ & \text{and } q = \left\lfloor \frac{N_y}{2} \right\rfloor + 2,\ldots,N_y, \\ e^{\frac{2\pi i(N_x-p+1)\mathrm{sgn}(\mu_1)h_1}{N_x}} e^{\frac{-2\pi i(q-1)\mathrm{sgn}(\mu_2)h_2}{N_y}}, & \text{for } p = \left\lfloor \frac{N_x}{2} \right\rfloor + 2,\ldots,N_x \\ & \text{and } q = 1,\ldots,\left\lfloor \frac{N_y}{2} \right\rfloor + 1, \\ e^{\frac{2\pi i(N_x-p+1)\mathrm{sgn}(\mu_1)h_1}{N_x}} e^{\frac{2\pi i(N_y-q+1)\mathrm{sgn}(\mu_2)h_2}{N_y}}, & \text{for } p = \left\lfloor \frac{N_x}{2} \right\rfloor + 2,\ldots,N_x \\ & \text{and } q = \left\lfloor \frac{N_y}{2} \right\rfloor + 2,\ldots,N_y, \end{cases} \tag{5.40}$$

as $h_1 = \frac{|\mu_1|\Delta t}{\Delta x} = |\mu_1|\Delta t N_x$ and $h_2 = \frac{|\mu_2|\Delta t}{\Delta y} = |\mu_2|\Delta t N_y$.

As with the MNIMC scheme for the 1D linear advection problem, it is important to investigate if this choice of eigenvalues, is real for the values of p and q that $\mathcal{F}_{p,q}(\mathbf{z})$ is real, for some $\mathbf{z} \in \mathbb{R}^{N_x N_y}$.

- When N_x and N_y are both even, $\mathcal{F}_{1,1}(\mathbf{z})$, $\mathcal{F}_{1,\frac{N_y}{2}+1}(\mathbf{z})$, $\mathcal{F}_{\frac{N_x}{2}+1,1}(\mathbf{z})$ and $\mathcal{F}_{\frac{N_x}{2}+1,\frac{N_y}{2}+1}(\mathbf{z})$ are all real. Consider,
$$\tilde{\lambda}_{1,\frac{N_y}{2}+1} = e^{-\pi i h_2}.$$
This is only real if $h_2 \in \mathbb{N}$. A similar problem exists when using (5.40) for $\tilde{\lambda}_{\frac{N_x}{2}+1,1}$ and $\tilde{\lambda}_{\frac{N_x}{2}+1,\frac{N_y}{2}+1}$.
- When N_x is even and N_y is odd, $\mathcal{F}_{1,1}(\mathbf{z})$ and $\mathcal{F}_{\frac{N_x}{2}+1,1}(\mathbf{z})$ are real. However using (5.40), $\tilde{\lambda}_{\frac{N_x}{2}+1,1}$ is only real if $h_1 \in \mathbb{N}$.
- When N_x is odd and N_y is even, $\mathcal{F}_{1,1}(\mathbf{z})$ and $\mathcal{F}_{1,\frac{N_y}{2}+1}(\mathbf{z})$ are real. However using (5.40), $\tilde{\lambda}_{1,\frac{N_y}{2}+1}$ is only real if $h_2 \in \mathbb{N}$.
- When N_x and N_y are both odd, $\mathcal{F}_{1,1}(\mathbf{z})$ is real. Using (5.40), $\tilde{\lambda}_{1,1} = 1$.

In order to investigate this problem for any values of $h_1, h_2 \in \mathbb{R}^+$, the problem will be restricted to N_x and N_y both odd. This ensures that the eigenvalues have the required conjugate pair properties.

It is interesting to note that these eigenvalues can be separated into two functions multiplied together, one in the x variable and the other in the y variable. These functions each take the form of the eigenvalues for the MNIMC scheme for the 1D linear advection problem. This is due to the independence of the solutions in the x- and y-directions. If we consider the 2D linear advection problem for a fixed $y \in [0, 1)$, the problem becomes the 1D linear advection problem in the x-direction. Similarly, if we fix $x \in [0, 1)$, the problem becomes the 1D linear advection problem in the y-direction.

Definition 5.2 (The 2D MNIMC scheme)**.** *Let Assumptions 5.1 hold true with $\mathbf{U}^n$ replaced by $\tilde{\mathbf{U}}^n$ to mark the difference in the schemes. Define the matrix $\tilde{M} \in \mathbb{R}^{N_x N_y \times N_x N_y}$ where N_x and N_y are both odd, by $\tilde{M} := V\tilde{\Lambda}V^*$, where the matrix V is defined as in Section 5.3.1 and $\tilde{\Lambda} \in \mathbb{C}^{N_x N_y \times N_x N y_y}$ contains the eigenvalues of the scheme in (5.40) along its main diagonal such that,*

$$\tilde{\Lambda}_{(q-1)N_x+p,(r-1)N_x+s} = \tilde{\lambda}_{p,q}\delta_{p,s}\delta_{q,r}, \tag{5.41}$$

for $p, s = 1, \ldots, N_x$ and $q, r = 1, \ldots, N_y$. The eigenvalues are positioned in the same order as the eigenvectors they correspond to in the matrix V. The scheme is implemented

by multiplying the vector containing the current state of the system by the matrix $\tilde{M}$*, as defined in Section 5.3 ie:* $\tilde{\mathbf{U}}^{n+1} = \tilde{M}\tilde{\mathbf{U}}^{n}$ *for all* $n \in \mathbb{N}_0$*.*

The eigenvalues of the MNIMC scheme for the 2D linear advection problem, all have unit magnitude. As a result the scheme is always numerically stable and numerically non-dissipative with respect to all wavenumber components of the numerical solution. The scheme is always numerically non-dispersive with respect to the resolvable wavenumber components of the numerical solution. When $h_1, h_2 \in \mathbb{N}$, the scheme is numerically non-dispersive with respect to all wavenumber component of the numerical solution by Section 5.4. When h_1 and/or h_2 is not in $\mathbb{N}$, the MNIMC scheme introduces aliasing errors into the numerical solution in the form of numerical dispersion. Before applying the MNIMC scheme to solve the 2D linear advection problem, we prove its consistency using similar methods to those employed in Lemma 3.9.

Lemma 5.3. *Suppose the initial condition* $u_0(x, y)$ *of problem (5.1) is a multiplicatively separable function such that* $u_0(x, y) = u_1(x)u_2(y)$*, where* $u_1, u_2 : \mathbb{R} \to \mathbb{R}$*,* $x \mapsto u_1(x)$ *and* $y \mapsto u_2(y)$ *have regularities* $r_1, r_2 \in \mathbb{N}_0$ *over* $(0, 1)$ *respectively. Also let Assumptions 5.1 hold true, allowing the MNIMC scheme to be defined as in Definition 5.2. Set the CFL number* $h \in \mathbb{R}^+$ *to be a fixed constant. Then the truncation error for the MNIMC scheme is such that,*

$$\tau_{j-1}^{n+1} = \mathcal{O}(\Delta x^{r_1}\Delta y^{r_2}) + \mathcal{O}(\Delta x^{r_1}) + \mathcal{O}(\Delta y^{r_2}). \tag{5.42}$$

Then for sufficiently smooth functions such that $r_1, r_2 \in \mathbb{N}$*,*

$$\tau_{j-1,k-1}^{n+1} \to 0 \text{ and } \Delta t \to 0 \text{ as } \Delta x, \Delta y \to 0,$$

for all $j = 1, \ldots, N_x$*,* $k = 1, \ldots, N_y$ *and* $n \in \mathbb{N}_0$*.*

Proof. The proof is identical to the proof of Lemma 3.9, but uses the result of Lemma 5.7 instead of Lemma 4.3. □

5.6.2 Implementing the MNIMC scheme for the 2D linear advection problem

In Section 5.6.1, the MNIMC scheme was defined for solving the 2D linear advection problem. Since the scheme was derived through the eigenvalues and eigenvectors of the scheme, we use the method of implementation for the scheme to derive its schematic

ie: $\tilde{\mathbf{U}}^{n+1} = \tilde{M}\tilde{\mathbf{U}}^{n}$ for all $n \in \mathbb{N}_0$. Formulating the schematic for the scheme in this way creates an explicit finite difference scheme defined by,

$$\begin{aligned}\tilde{U}_{j,k}^{n+1} &= \frac{1}{N_x N_y}\sum_{p=0}^{N_x-1}\sum_{q=0}^{N_y-1}\left\{1+2\sum_{r=1}^{\frac{N_x-1}{2}}\cos\left[\frac{2\pi r(j-p-\operatorname{sgn}(\mu_1)h_1)}{N_x}\right]\right\}\{1 \\ &\quad +2\sum_{s=1}^{\frac{N_y-1}{2}}\cos\left[\frac{2\pi s(k-q-\operatorname{sgn}(\mu_2)h_2)}{N_y}\right]\Bigg\}\tilde{U}_{p,q}^{n}, \qquad (5.43)\end{aligned}$$

for $j, p = 0, \ldots, N_x - 1$ and $k, q = 0, \ldots, N_y - 1$. This schematic shows that the current state of the system at every grid point in space is utilised to calculate the state of the system Δt in time, at each grid point in space. This potentially means that the matrix $\tilde{M}$ is a full matrix, so the scheme will be computationally expensive to implement.

Figure 5.2 shows the results of applying the MNIMC for the 2D linear advection problem to the 2D square function initial condition defined in (5.134), using $h_1 = h_2 = \frac{1}{2}$. We can see a shifted periodic nature in the results of the scheme, similar to that seen in MNIMC scheme for the 1D linear advection problem. At odd multiples of Δt, oscillations are present in the numerical solution, whilst at even multiples of Δt, the 2D square function is recovered. As aliasing is the only error introduced into the numerical solution by the scheme, it must be aliasing causing these oscillations. As with the MNIMC scheme for the 1D linear advection problem, the shifted $2\Delta t$-periodic nature seen here, is likely due to the denominators of h_1 and h_2 being equal to two. Aliasing errors in the MNIMC scheme for the 2D linear advection problem are investigated in Lemma 5.6 of Section 5.8.

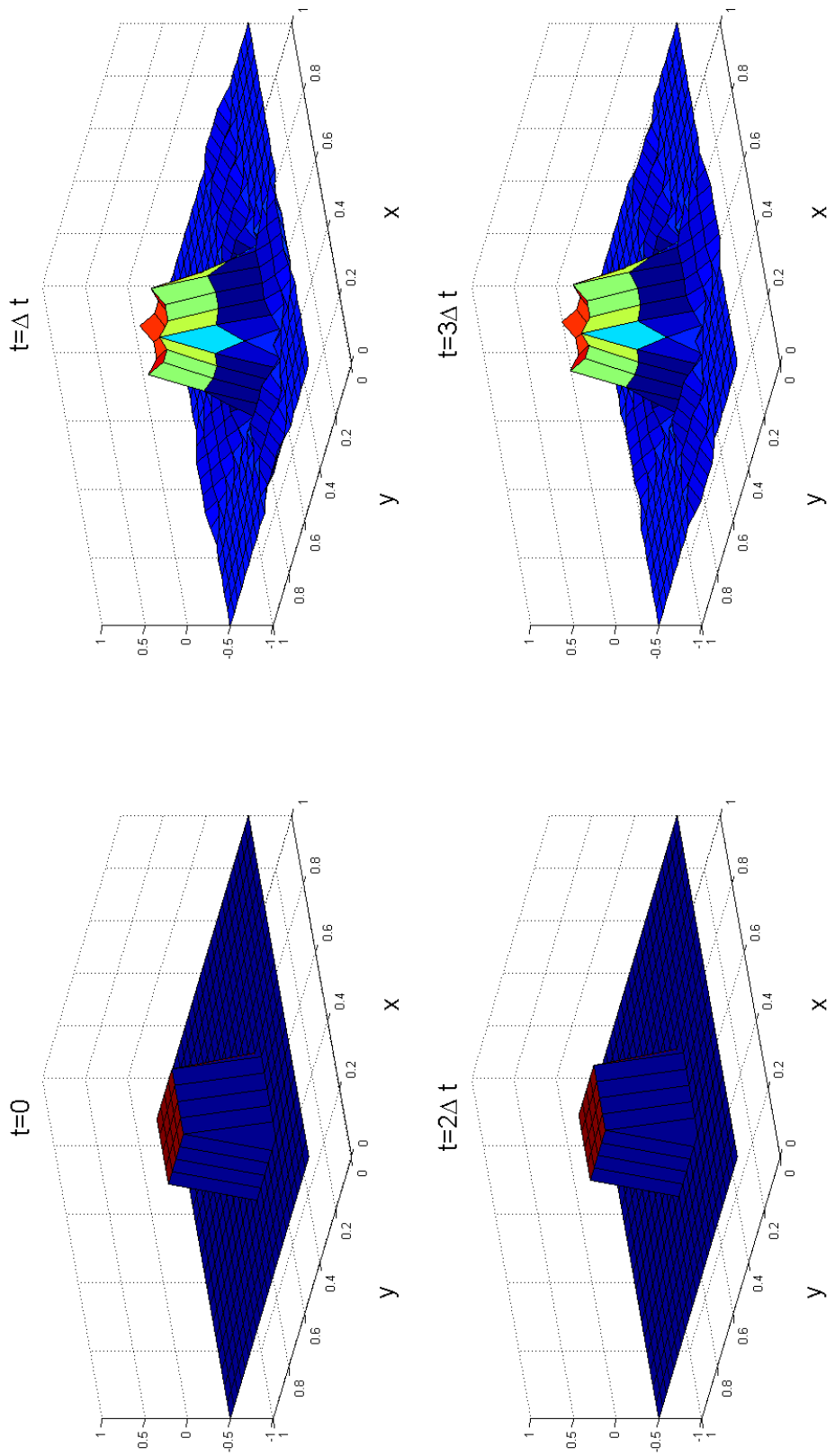


Figure 5.2: *The numerical results from applying the MNIMC scheme, for the 2D linear advection problem, to the 2D square function initial condition in (5.134). The effects of aliasing errors in the scheme can be seen every* $2\Delta t$*. These results were generated using* $N_x = N_y = 21$*,* $\mu_1 = \mu_2 = 1$ *and* $h_1 = h_2 = \frac{1}{2}$ $\left(\Delta t = \frac{1}{42}\right)$*.*

5.7 Dissipative and dispersive metrics

The dissipative and dispersive metrics for the 2D linear advection problem, are defined by easily extending the definition for the metrics of the 1D problem in Section 3.8. The numerically dissipative and dispersive properties of the schemes for the 2D linear advection problem, are determined by h_1 and h_2, therefore the metrics will be functions of these variables. By defining these metrics for the 2D linear advection problem, we are able to verify our analysis of the dissipative and dispersive properties of the Upwind and Crank-Nicolson schemes in Table 5.1.

5.7.1 The dissipative metric

Definition 5.4 (Dissipative Metric)**.** *Define two finite difference schemes for solving problem (5.1), using the same spatial step size $\Delta x > 0$ and $\Delta y > 0$, and temporal step size $\Delta t > 0$. Let the eigenvalues of the scheme we wish to find the metric for, be denoted by $\left\{\lambda^{(1)}_{p,q}(h_1,h_2)\right\}^{N_x,N_y}_{p=1,q=1}$. Also, let the eigenvalues of the reference scheme be denoted by $\left\{\lambda^{(2)}_{p,q}(h_1,h_2)\right\}^{N_x,N_y}_{p=1,q=1}$. Let the (p,q)th eigenvalue of each scheme correspond to the (p,q)th eigenvector of the 2D DFT basis, as defined in Section 5.3. Define the vectors $\mathbf{z}_1(h_1,h_2), \mathbf{z}_2(h_1,h_2) \in \mathbb{R}^{N_xN_y}$ such that $[\mathbf{z}_j(h_1,h_2)]_{p,q} = |\lambda^{(j)}_{p,q}(h_1,h_2)|^2$ for $p = 1,\ldots,N_x$, $q = 1,\ldots,N_y$ and $j = 1,2$. Then the numerically dissipative metric is defined by $d_{dissipative} : \mathbb{R}^{N_xN_y} \times \mathbb{R}^{N_xN_y} \to \mathbb{R}$, such that,*

$$
\begin{aligned}
& d_{dissipative}(\mathbf{z}_1(h_1,h_2),\mathbf{z}_2(h_1,h_2)) \\
= \; & \frac{1}{2\left\lfloor\frac{N_x}{2}\right\rfloor\left\lfloor\frac{N_y}{2}\right\rfloor + \left\lfloor\frac{N_x}{2}\right\rfloor + \left\lfloor\frac{N_y}{2}\right\rfloor + 1}\left[\sum_{p=1}^{\left\lfloor\frac{N_x}{2}\right\rfloor+1}\sum_{q=1}^{\left\lfloor\frac{N_y}{2}\right\rfloor+1} |[\mathbf{z}_1(h_1,h_2)]_{p,q} - [\mathbf{z}_2(h_1,h_2)]_{p,q}|\right. \\
& \left. + \sum_{p=2}^{\left\lfloor\frac{N_x}{2}\right\rfloor+1}\sum_{q=\left\lfloor\frac{N_y}{2}\right\rfloor+2}^{N_y} |[\mathbf{z}_1(h_1,h_2)]_{p,q} - [\mathbf{z}_2(h_1,h_2)]_{p,q}|\right] \\
= \; & \frac{1}{2\left\lfloor\frac{N_x}{2}\right\rfloor\left\lfloor\frac{N_y}{2}\right\rfloor + \left\lfloor\frac{N_x}{2}\right\rfloor + \left\lfloor\frac{N_y}{2}\right\rfloor + 1}\left[\sum_{p=1}^{\left\lfloor\frac{N_x}{2}\right\rfloor+1}\sum_{q=1}^{\left\lfloor\frac{N_y}{2}\right\rfloor+1} \left||\lambda^{(1)}_{p,q}(h_1,h_2)|^2 - |\lambda^{(2)}_{p,q}(h_1,h_2)|^2\right|\right. \\
& \left. + \sum_{p=2}^{\left\lfloor\frac{N_x}{2}\right\rfloor+1}\sum_{q=\left\lfloor\frac{N_y}{2}\right\rfloor+2}^{N_y} \left||\lambda^{(1)}_{p,q}(h)|^2 - |\lambda^{(2)}_{p,q}(h)|^2\right|\right]
\end{aligned}
\tag{5.44}
$$

In the case of the 2D linear advection problem, the scheme with eigenvalues which

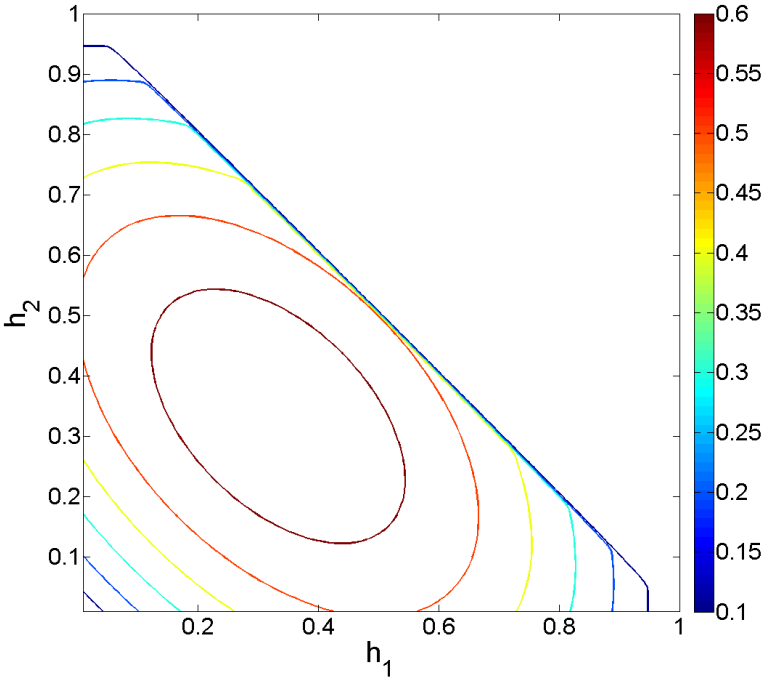
are numerically non-dissipative with respect to the resolvable wavenumber components, is the MNIMC scheme in Section 5.6.1. We can then use this scheme as our reference scheme in the dissipative metric for the Upwind and Crank-Nicolson schemes, for the 2D linear advection problem. Using the fact that N_x and N_y are both required to be odd for the MNIMC scheme, this results in,

$$d_{dissipative}(h_1, h_2) = \frac{2}{N_x N_y + 1} \left[\sum_{p=1}^{\frac{N_x+1}{2}} \sum_{q=1}^{\frac{N_y+1}{2}} \left| |\lambda_{p,q}(h_1, h_2)|^2 - 1 \right| + \sum_{p=2}^{\frac{N_x+1}{2}} \sum_{\frac{N_y+3}{2}}^{N_y} \left| |\lambda_{p,q}(h_1, h_2)|^2 - 1 \right| \right] \tag{5.45}$$

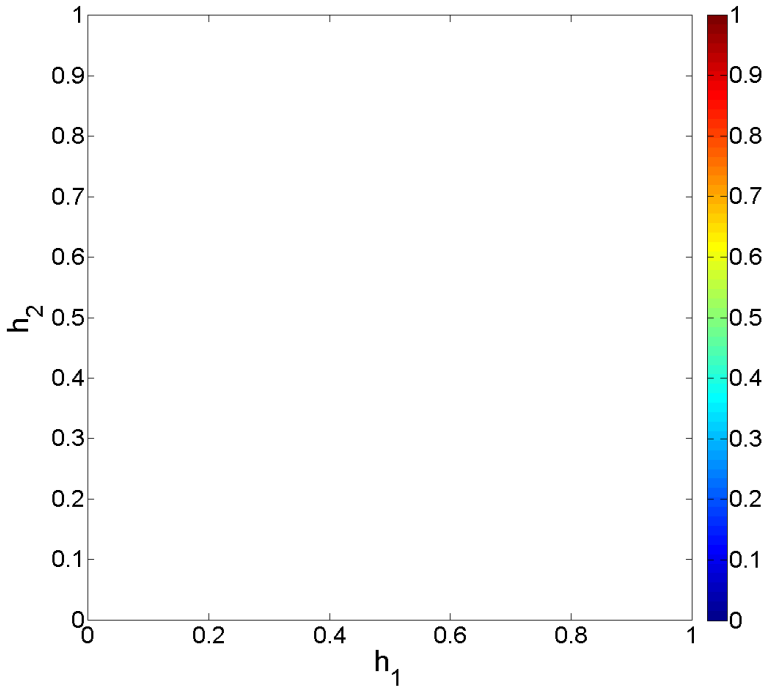
Figure 5.3 displays the results of the dissipative metric for the Upwind and Crank-Nicolson schemes, when the Upwind scheme is numerically stable ie: $h_1 + h_2 = h \leq 1$. In this region, the dissipative metric in (5.45) is less than or equal to one due to the stability of the schemes. Figure 5.3(a) shows that the dissipative metric for the Upwind scheme is never zero, so the scheme is never non-dissipative with respect to the resolvable wavenumber components of the numerical solution. This agrees with the analysis of the scheme in Table 5.1. However the metric tends towards zero when $h_1, h_2 \to 0^+$, or when $h_1 = 1$ and $h_2 \to 0^+$ and finally as $h_1 \to 0^+$ when $h_2 = 1$. These limits correspond to when the Upwind scheme for the 1D linear advection problem is numerically non-dissipative with respect to the resolvable wavenumber components of the numerical solution.

The blue line along $h_1 + h_2 = 1$, is due to the sharp gradient between the metric and the zeros plotted for $h_1 + h_2 > 1$. It is not an indication that the metric is zero. In order to prevent this gradient from being mistaken for the metric equalling zero, it would be best to view the results of the metrics as surface plots rather than contour plots. Despite this, we present contour plots in Figures 5.3 and 5.4, as they are easier to present results in this thesis. Another solution is to plot the metric for $(h_1, h_2) \in (0, 1] \times (0, 1]$, with the knowledge that the Upwind scheme will be unstable for $h_1 + h_2 > 1$.

Figure 5.3(b) depicts the dissipative metric for the Crank-Nicolson scheme. The plot does not appear to show anything as the dissipative metric is zero for any value of h_1 and h_2. This agrees with the analysis of the scheme in Table 5.1, which showed that the scheme is always numerically non-dissipative with respect to all wavenumber components of the numerical solution.



(a) The dissipative metric for the 2D Upwind scheme.



(b) The dissipative metric for the 2D Crank-Nicolson scheme.

Figure 5.3: *The dissipative metric in Equation (5.45) applied to the Upwind and Crank-Nicolson schemes using* $\mu_1 = \mu_2 = 1$, $N_x = 101$, $N_y = 51$ *and considering* $h_1 + h_2 = h \leq 1$.

5.7.2 The dispersive metric

Definition 5.5 (Dispersive Metric)**.** *Define two finite difference schemes for solving problem (5.1), using the same spatial step size* $\Delta x > 0$ *and* $\Delta y > 0$*, and temporal step*

size $\Delta t > 0$. Let the eigenvalues of the scheme we wish to find the metric for, be denoted by $\left\{\lambda^{(1)}_{p,q}(h_1,h_2)\right\}^{N_x,N_y}_{p=1,q=1}$. Also, let the eigenvalues of the reference scheme be denoted by $\left\{\lambda^{(2)}_{p,q}(h_1,h_2)\right\}^{N_x,N_y}_{p=1,q=1}$. Let the (p,q)th eigenvalue of each scheme correspond to the (p,q)th eigenvector of the 2D DFT basis, as defined in Section 5.3. Define the vectors $\mathbf{z}_1(h_1,h_2), \mathbf{z}_2(h_1,h_2) \in \mathbb{R}^{N_xN_y}$ such that $[\mathbf{z}_j(h_1,h_2)]_{p,q} = \theta^{(j)}_{p,q}(h_1,h_2)$ for $= 1,\ldots,N_x$, $q = 1,\ldots,N_y$ and $j = 1,2$. Then the numerically dispersive metric is defined by $d_{dispersive} : \mathbb{C}^{N_xN_y} \times \mathbb{C}^{N_xN_y} \to \mathbb{R}$, such that,

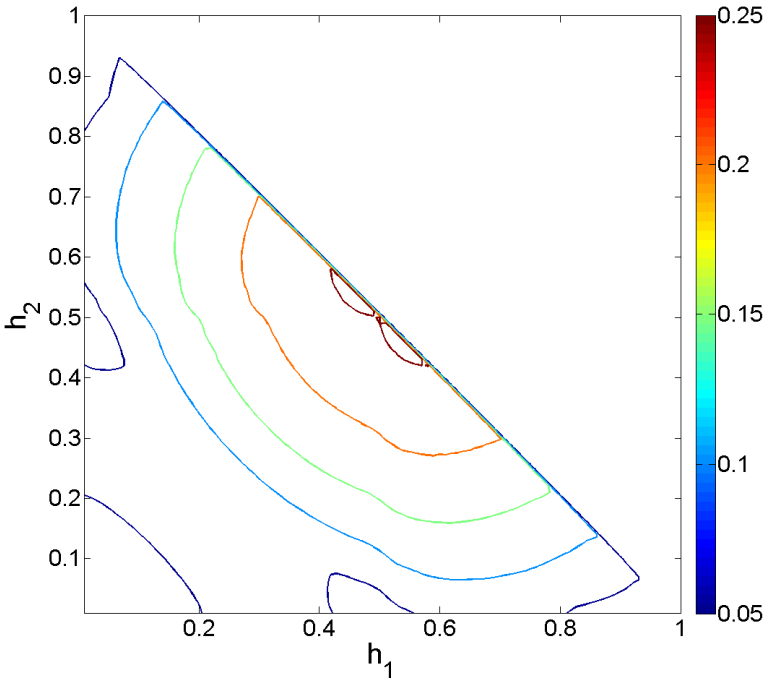
$$
\begin{aligned}
& d_{dispersive}(\mathbf{z}_1(h_1,h_2),\mathbf{z}_2(h_1,h_2)) \\
= \; & \frac{1}{2\pi\left(2\left\lfloor\frac{N_x}{2}\right\rfloor\left\lfloor\frac{N_y}{2}\right\rfloor + \left\lfloor\frac{N_x}{2}\right\rfloor + \left\lfloor\frac{N_y}{2}\right\rfloor + 1\right)}\left[\sum_{p=1}^{\left\lfloor\frac{N_x}{2}\right\rfloor+1}\sum_{q=1}^{\left\lfloor\frac{N_y}{2}\right\rfloor+1} |[\mathbf{z}_1(h_1,h_2)]_{p,q}\right. \\
& \left. -[\mathbf{z}_2(h_1,h_2)]_{p,q}| + \sum_{p=2}^{\left\lfloor\frac{N_x}{2}\right\rfloor+1}\sum_{\left\lfloor\frac{N_y}{2}\right\rfloor+2}^{N_y} |[\mathbf{z}_1(h_1,h_2)]_{p,q} - [\mathbf{z}_2(h_1,h_2)]_{p,q}|\right] \\
= \; & \frac{1}{2\pi\left(2\left\lfloor\frac{N_x}{2}\right\rfloor\left\lfloor\frac{N_y}{2}\right\rfloor + \left\lfloor\frac{N_x}{2}\right\rfloor + \left\lfloor\frac{N_y}{2}\right\rfloor + 1\right)}\left[\sum_{p=1}^{\left\lfloor\frac{N_x}{2}\right\rfloor+1}\sum_{q=1}^{\left\lfloor\frac{N_y}{2}\right\rfloor+1} \left|\theta^{(1)}_{p,q}(h_1,h_2)\right.\right. \\
& \left.\left. -\theta^{(2)}_{p,q}(h_1,h_2)\right| + \sum_{p=2}^{\left\lfloor\frac{N_x}{2}\right\rfloor+1}\sum_{\left\lfloor\frac{N_y}{2}\right\rfloor+2}^{N_y} \left|\theta^{(1)}_{p,q}(h_1,h_2) - \theta^{(2)}_{p,q}(h_1,h_2)\right|\right]
\end{aligned}
\tag{5.46}
$$

In the case of the 2D linear advection problem, the scheme with eigenvalues which are numerically non-dispersive with respect to the resolvable wavenumber components of the numerical solution, is the MNIMC scheme in Section 5.6.1. We can use this scheme as our reference scheme in the dispersive metric for the Upwind and Crank-Nicolson schemes, for the 2D linear advection problem. Using the fact that N_x and N_y are both required to be odd for the MNIMC scheme, this results in,

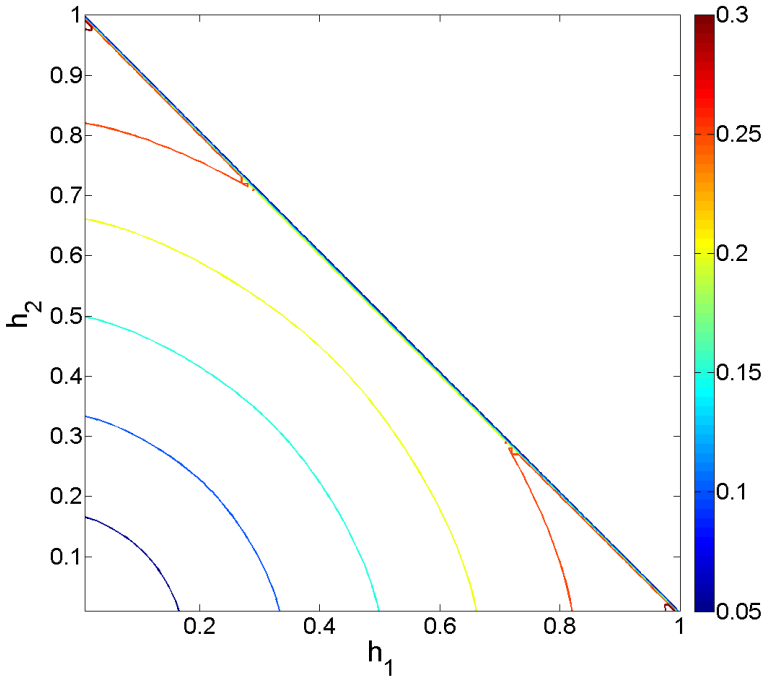
$$
\begin{aligned}
& d_{dispersive}(h_1,h_2) \\
= \; & \frac{1}{\pi(N_xN_y+1)}\sum_{p=1}^{\frac{N_x+1}{2}}\sum_{q=1}^{\frac{N_y+1}{2}} \left|\theta_{p,q}(h_1,h_2) - \left(-\frac{2\pi(p-1)h_1}{N_x} - \frac{2\pi(q-1)h_2}{N_y}\right)\right| \\
& + \sum_{p=2}^{\frac{N_x+1}{2}}\sum_{\frac{N_y+3}{2}}^{N_y} \left|\theta_{p,q}(h_1,h_2) - \left(\frac{-2\pi(p-1)h_1}{N_x} + \frac{2\pi(N_y-q+1)h_2}{N_y}\right)\right|
\end{aligned}
\tag{5.47}
$$

As before, we consider the dispersive metric in (5.47) for $h_1 + h_2 = h \leq 1$, where the

Upwind and Crank-Nicolson schemes are both numerically stable. This results in the dispersive metric being less than or equal to one.



(a) The dispersive metric for the 2D Upwind scheme.



(b) The dispersive metric for the 2D Crank-Nicolson scheme.

Figure 5.4: *The dispersive metric in Equation (5.47), applied to the Upwind and Crank-Nicolson schemes for the 2D linear advection problem, using* $\mu_1 = \mu_2 = 1$, $N_x = 101$, $N_y = 51$ *and considering* $h_1 + h_2 = h \leq 1$.

Figure 5.4(a) shows that the Upwind scheme is always numerically dispersive as the

metric is never zero. The metric tends to zero when, $h_1 \to 0^+$ and h_2 is equal to 1 or $\frac{1}{2}$ or as $h_2 \to 0^+$. Similarly, as $h_2 \to 0^+$ when h_1 is equal to 1 or $\frac{1}{2}$ or as $h_1 \to 0^+$. The metric tends towards zero at these points as letting either h_1 or h_2 equal zero, reduces the dimension of the system and we recover the conditions under which the Upwind scheme for the 1D problem, is numerically non-dispersive. We again obtain a misleading indication in this Figure that the metric is zero along the line $h_1 + h_2 = 1$. This is due to the sharp gradient in the plot as discussed in the previous Section, for the dissipative metric for the Upwind scheme.

Figure 5.4(b) shows that the Crank-Nicolson scheme is always numerically dispersive. As $h_1, h_2 \to 0^+$, the numerically dispersive properties of the scheme decrease.

5.8 Aliasing error in the MNIMC scheme

Similarly to the 1D linear advection problem, we wish to use the MNIMC scheme for the 2D linear advection problem, to construct perfect observations of the system. As for the 1D linear advection problem, we want to construct these perfect observations using the MNIMC scheme for the 2D linear advection problem, but to do this, we need to define the global error in the MNIMC scheme for the 2D linear advection problem. Let $\tilde{\mathbf{x}}_0 \in \mathbb{R}^{N_x N_y}$ denote the true initial condition $u_0(x, y)$, sampled at the spatial grid points defined in Assumptions 5.1, such that $\{\tilde{\mathbf{x}}_0\}_{(k-1)N_x+j} := u_0(x_{j-1}, y_{k-1})$. Now define $\tilde{\mathbf{x}}_l \in \mathbb{R}^{N_x N_y}$ by $\tilde{\mathbf{x}}_l := \tilde{M}^l \tilde{\mathbf{x}}_0$ for all $l \in \mathbb{N}$. Then the global error in the MNIMC scheme $\mathbf{r}_l \in \mathbb{R}^{N_x N_y}$, is defined by,

$$\mathbf{r}_l := \tilde{\mathbf{y}} - \tilde{M}^l \tilde{\mathbf{x}}_0, \tag{5.48}$$

where $\mathbf{y} := \tilde{\mathbf{y}}_l$ denotes the lth set of perfect observations such that $\{\tilde{\mathbf{y}}_l\}_{(k-1)N_x+j} := u(x_{j-1}, y_{k-1}, l\Delta t)$ for $j = 1, \ldots, N_x$ and $k = 1, \ldots, N_y$. As only aliasing errors are introduced by the MNIMC scheme, $\mathbf{r}_l$ can again be viewed as an additive correction term to correct for aliasing errors in $\tilde{M}^l \tilde{\mathbf{x}}_0$ such that,

$$\tilde{\mathbf{y}}_l = \tilde{\mathbf{x}}_l + \mathbf{r}_l = \tilde{M}^l \tilde{\mathbf{x}}_0 + \mathbf{r}_l. \tag{5.49}$$

Since $\tilde{\mathbf{x}}_0$ is defined as the discrete sample of $u_0(x, y)$, $\mathbf{r}_0 = \mathbf{0}$. Choosing $h_1 = h_2 = 1$, results in $\tilde{M} = M_{NIMC}$, where $h_{NIMC} = 1$, so $\mathbf{r}_l = \mathbf{0}$ for all $l \in \mathbb{N}_0$. In order to use (5.49) in our strong constraint 4D-Var problem, we need to analyse the properties of $\mathbf{r}_l$. The following Lemma investigates the shifted periodic nature of the aliasing error present in the MNIMC scheme for solving the 2D linear advection problem.

Lemma 5.6. *Let Assumptions 5.1 hold true, allowing the MNIMC scheme do be defined as in Definition 5.2. Also, let $u_0(x, y)$ possess a convergent Fourier series.*

Additionally, let the CFL numbers in the x- and y-directions be expressed as $h_1 = \frac{q_1}{b_1}$ *and* $h_2 = \frac{q_2}{b_2}$*, such that* $q_j, b_j \in \mathbb{Z}$ *and* $gcd(q_j, b_j) = 1$ *for* $j = 1, 2$*. Also, let* $c = lcm(b_1, b_2)$ *(lowest common multiple). Then the aliasing error in* $\tilde{\mathbf{x}}_l$*, generated by* $\tilde{M}$*, denoted by* $\mathbf{r}_l$ *is such that,*

$$\mathbf{r}_l = \begin{cases} \mathbf{0}, & \text{for } [l]_c = 0, \\ \tilde{M}^{l-[l]_c}\mathbf{r}_{[l]_c}, & \text{for } [l]_c = 1, \ldots, c-1, \end{cases} \tag{5.50}$$

for all l*, where* $[\cdot]_c$ *denotes modulo* c*.*

Proof. This proof follows the same method as the proof of Lemma 3.12. As we are considering the MNIMC scheme for solving the 2D linear advection problem, we are restricted to N_x and N_y both odd. Rearranging (5.49) and applying the 2D DFT results in,

$$\mathcal{F}_{p,q}(\mathbf{r}_l) = \mathcal{F}_{p,q}(\tilde{\mathbf{y}}_l) - \tilde{\lambda}_{p,q}^l \mathcal{F}_{p,q}(\tilde{\mathbf{x}}_0), \tag{5.51}$$

for all $l \in \mathbb{N}_0$. The vector $\tilde{\mathbf{y}}_l$ contains a discrete sample of the true physical system sampled at each grid point in space, at time $l\Delta t$,

$$[\tilde{\mathbf{y}}_l]_{(k-1)N_x+j} = u(x_j, y_k, l\Delta t) = u(x_{j-1} - \mu_1 l\Delta t, y_{k-1} - \mu_2 l\Delta t, 0), \tag{5.52}$$

for all $j = 1, \ldots, N_x$ and $k = 1, \ldots, N_y$. As with Lemma 3.12, we would like to use the Fourier series of $u_0(x, y)$ in (5.5), to represent $u(x - \mu_1 l\Delta t, y - \mu_2 l\Delta t, 0)$. The function $u(x, y, 0)$ is the periodic extension of the function $u_0(x, y)$. Under the conditions of the Lemma, this Fourier series is convergent. If $u_0(x, y)$ is continuous over $[0, 1) \times [0, 1)$, $\lim_{x\to 0^+} u_0(x, y) = \lim_{x\to 1^-} u_0(x, y)$ for all $y \in [0, 1)$, $\lim_{y\to 0^+} u_0(x, y) = \lim_{y\to 1^-} u_0(x, y)$ for all $x \in [0, 1)$ and

$$\lim_{x\to 0^+} \lim_{y\to 0^+} u_0(x, y) = \lim_{x\to 1^-} \lim_{y\to 1^-} u_0(x, y),$$

then the Fourier series is equal to the function $u(x, y, 0)$ for all $(x, y) \in \mathbb{R} \times \mathbb{R}$. In this instance, we can use the Fourier series of $u_0(x, y)$ to represent $u(x_{p-1} - \mu_1 l\Delta t, y_{q-1} - \mu_2 l\Delta t, 0)$ for all $p = 1, \ldots, N_x$ and $q = 1, \ldots, N_y$.

However, if $u_0(x, y)$ is piecewise continuous over $[0, 1) \times [0, 1)$, there exists $x \in [0, 1)$ such that $\lim_{y\to 0^+} u_0(x, y) \neq \lim_{y\to 1^-} u_0(x, y)$, there exists $y \in [0, 1)$ such that $\lim_{x\to 0^+} u_0(x, y) \neq \lim_{x\to 1^-} u_0(x, y)$ or

$$\lim_{x\to 0^+} \lim_{y\to 0^+} u_0(x, y) \neq \lim_{x\to 1^-} \lim_{y\to 1^-} u_0(x, y),$$

then the function possesses discontinuities. We assume that these are jump discontinuities by the conditions of the Lemma as discussed in Section 5.2. Then as the Fourier series is convergent, the Fourier series is equal to $u_0(x, y)$ at every continuous point and not equal to $u_0(x, y)$ at each point of discontinuity. When the sample

points of $u_0(x,y)$ do not sample a point of discontinuity, the Fourier series of $u_0(x,y)$ can be used to represent $u(x_{p-1}-\mu_1 l\Delta t, y_{q-1}-\mu_2 l\Delta t, 0)$ for all $p=1,\ldots,N_x$ and $q=1,\ldots,N_y$ as in the continuous case. When for a given l, the (p,q)th sample point coincides with a discontinuity in the function, the Fourier series of $u_0(x,y)$ is not equal to $u(x_{p-1}-\mu_1 l\Delta t, y_{q-1}-\mu_2 l\Delta t, 0)$, so cannot be used to represent this sample point.

As before we take advantage of the non-uniqueness of the 2D DFT [66]. We define a new function based on $u(x,y,0)$, that has a convergent Fourier series and is continuous and equal to $u(x,y,0)$ at every sample point. The Fourier series of this function will then be equal to $u(x,y,0)$ at all of the sample points on the domain. Define,

$$\hat{X}=\left\{(x,y)\in[0,1)\times[0,1)\;\middle|\;\begin{array}{l}u(x,y,0)\text{ is a point discontinuity in the domain or}\\ \text{is a point along a line of discontinuity forming}\\ \text{a boundary along a discontinuous pieces}\\ \text{of the domain.}\end{array}\right\} \tag{5.53}$$

The jump discontinuities in $u(x,y,0)$ over $[0,1)\times[0,1)$ are either point discontinuities or form lines in the (x,y)-domain, along the boundaries of the discontinuous pieces of $u(x,y,0)$.

Now define $\hat{X}_l\subseteq\hat{X}$ for each $l\in\mathbb{N}_0$,

$$\hat{X}_l=\left\{(\hat{x},\hat{y})\in\hat{X}\;\middle|\;\begin{array}{l}\exists p\in\{1,\ldots,N_x\}\text{ and }\exists q\in\{1,\ldots,N_y\}\text{ such that}\\ \hat{x}=[x_{p-1}-\mu_1 l\Delta t]_1\text{ and }\hat{y}=[y_{q-1}-\mu_2 l\Delta t]_1\end{array}\right\}. \tag{5.54}$$

This set identifies the sample points within $[0,1)\times[0,1)$ where discontinuities lie within $u(x-\mu_1 l\Delta t, y-\mu_2 l\Delta t,0)$. If there exists l such that $u(x_{p-1}-\mu_1 l\Delta t, x_{q-1}-\mu_2 l\Delta t,0)$ is a continuous point for all $p=1,\ldots,N_x$ and $q=1,\ldots,N_y$, then $\hat{X}_l=\emptyset$. When $u(x,y,0)$ is continuous over $\mathbb{R}\times\mathbb{R}\times\{0\}$, $\hat{X}=\emptyset$, hence $\hat{X}_l=\emptyset$ for all $l\in\mathbb{N}_0$.

Consider $(\hat{x},\hat{y})\in\hat{X}_l$, then there exists $p\in\{1,\ldots,N_x\}$ and $q\in\{1,\ldots,N_y\}$ such that, $\hat{x}=[x_{p-1}-\mu_1 l\Delta t]_1$ and $\hat{y}=[y_{q-1}-\mu_2 l\Delta t]_1$. Then by (3.60) in the proof of Lemma 3.12, there exists $p\in\{1,\ldots,N_x\}$ and $q\in\{1,\ldots,N_y\}$ such that,

$$\hat{x} = [x_{[p-1-\mathrm{sgn}(\mu_1)h_1(l-[l]_{b_1})]_{N_x}}-\mu_1[l]_{b_1}\Delta t]_1,\ \text{ using } h_1=\frac{|\mu_1|\Delta t}{\Delta x}, \tag{5.55}$$

$$\hat{y} = [y_{[q-1-\mathrm{sgn}(\mu_2)h_2(l-[l]_{b_2})]_{N_y}}-\mu_2[l]_{b_2}\Delta t]_1,\ \text{ using } h_2=\frac{|\mu_2|\Delta t}{\Delta y}. \tag{5.56}$$

Then as $[p-1-\mathrm{sgn}(\mu_1)h_1(l-1[l]_{b_1})]_{N_x}\in\{0,\ldots,N_x-1\}$, there exists some $s_1\in\{1,\ldots,N_x\}$ such that $s_1-1=[p-1-\mathrm{sgn}(\mu_1)h_1(l-[l]_{b_1})]_{N_x}$, hence $\hat{x}=[x_{s_1-1}-\mu_1[l]_{b_1}\Delta t]_1$. Similarly, there exists $s_2\in\{1,\ldots,N_y\}$ such that $\hat{y}=[y_{s_2-1}-\mu_2[l]_{b_2}\Delta t]_1$. This means that $(\hat{x},\hat{y})\in\hat{X}_l\Leftrightarrow(\hat{x},\hat{y})\in\hat{X}_{\mathrm{lcm}(b_1,b_2)}$. Therefore, $\hat{X}_l=\hat{X}_{[l]_{lcm(b_1,b_2)}}$ for all $l\in\mathbb{N}_0$. As $\hat{X}_l=\hat{X}_{[l]_{\mathrm{lcm}(b_1,b_2)}}$, we have shown that there are at most $\mathrm{lcm}(b_1,b_2)$ different subsets of points in $u(x,y,0)$ over $[0,1)\times[0,1)$, where a discontinuity is sampled, over time.

We will now define our new functions, $v_l : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, such that $(x, y) \mapsto v_l(x, y)$, for $l = 0, \ldots, \mathrm{lcm}(b_1, b_2) - 1$ such that $\hat{X}_l \neq \emptyset$, using the points $\hat{X}_l = \{(\hat{x}_j, \hat{y}_j)\}_{j=1}^{|\hat{X}_l|}$,

$$
v_l(x,y) = \begin{cases}
\left[\frac{u\left(\hat{x}_j-\frac{\Delta x}{2},\hat{y}_j-\frac{\Delta y}{2},0\right)+u\left(\hat{x}_j-\frac{\Delta x}{2},\hat{y}_j+\frac{\Delta y}{2},0\right)-2u(\hat{x}_j,\hat{y}_j,0)}{\Delta x}\right](\hat{x}_j-x) & \text{for } x\in\left[\hat{x}_j-\frac{\Delta x}{2},\hat{x}_j\right] \\
+\left[\frac{u\left(\hat{x}_j-\frac{\Delta x}{2},\hat{y}_j-\frac{\Delta y}{2},0\right)-u(\hat{x}_j-\frac{\Delta x}{2},\hat{y}_j+\frac{\Delta y}{2},0)}{\Delta y}\right](\hat{y}_j-y) & \text{and } y\in\left[\frac{\Delta y}{\Delta x}(x-\hat{x}_j)+\hat{y}_j,\frac{\Delta y}{\Delta x}(\hat{x}_j-x)+\hat{y}_j\right), \\
+u(\hat{x}_j,\hat{y}_j,0), & \text{where } (\hat{x}_j,\hat{y}_j)\in\hat{X}_l, \\
\left[\frac{u\left(\hat{x}_j-\frac{\Delta x}{2},\hat{y}_j-\frac{\Delta y}{2},0\right)-u(\hat{x}_j+\frac{\Delta x}{2},\hat{y}_j-\frac{\Delta y}{2},0)}{\Delta x}\right](\hat{x}_j-x) & \text{for } x\in\left(\frac{\Delta x}{\Delta y}(y-\hat{y}_j)+\hat{x}_j,\frac{\Delta x}{\Delta y}(\hat{y}_j-y)+\hat{x}_j\right) \\
+\left[\frac{u\left(\hat{x}_j+\frac{\Delta x}{2},\hat{y}_j-\frac{\Delta y}{2},0\right)+u\left(\hat{x}_j-\frac{\Delta x}{2},\hat{y}_j-\frac{\Delta y}{2},0\right)-2u(\hat{x}_j,\hat{y}_j,0)}{\Delta y}\right](\hat{y}_j-y) & \text{and } y\in\left[\hat{y}_j-\frac{\Delta y}{2},\hat{y}_j\right), \\
+u(\hat{x}_j,\hat{y}_j,0), & \text{where } (\hat{x}_j,\hat{y}_j)\in\hat{X}_l, \\
-\left[\frac{u\left(\hat{x}_j+\frac{\Delta x}{2},\hat{y}_j+\frac{\Delta y}{2},0\right)+u(\hat{x}_j+\frac{\Delta x}{2},\hat{y}_j-\frac{\Delta y}{2},0)-2u(\hat{x}_j,\hat{y}_j,0)}{\Delta x}\right](\hat{x}_j-x) & \text{for } x\in\left(\hat{x}_j,\hat{x}_j+\frac{\Delta x}{2}\right) \\
+\left[\frac{u\left(\hat{x}_j+\frac{\Delta x}{2},\hat{y}_j-\frac{\Delta y}{2},0\right)-u\left(\hat{x}_j+\frac{\Delta x}{2},\hat{y}_j+\frac{\Delta y}{2},0\right)}{\Delta y}\right](\hat{y}_j-y) & \text{and } y\in\left[\frac{\Delta y}{\Delta x}(\hat{x}_j-x)+\hat{y}_j,\frac{\Delta y}{\Delta x}(x-\hat{x}_j)+\hat{y}_j\right), \\
+u(\hat{x}_j,\hat{y}_j,0), & \text{where } (\hat{x}_j,\hat{y}_j)\in\hat{X}_l, \\
\left[\frac{u\left(\hat{x}_j-\frac{\Delta x}{2},\hat{y}_j+\frac{\Delta y}{2},0\right)-u(\hat{x}_j+\frac{\Delta x}{2},\hat{y}_j+\frac{\Delta y}{2},0)}{\Delta x}\right](\hat{x}_j-x) & \text{for } x\in\left[\frac{\Delta x}{\Delta y}(\hat{y}_j-y)+\hat{x}_j,\frac{\Delta x}{\Delta y}(y-\hat{y}_j)+\hat{x}_j\right) \\
-\left[\frac{u\left(\hat{x}_j+\frac{\Delta x}{2},\hat{y}_j+\frac{\Delta y}{2},0\right)+u\left(\hat{x}_j-\frac{\Delta x}{2},\hat{y}_j+\frac{\Delta y}{2},0\right)-2u(\hat{x}_j,\hat{y}_j,0)}{\Delta y}\right](\hat{y}_j-y) & \text{and } y\in\left(\hat{y}_j,\hat{y}_j+\frac{\Delta y}{2}\right), \\
+u(\hat{x}_j,\hat{y}_j,0), & \text{where } (\hat{x}_j,\hat{y}_j)\in\hat{X}_l, \\
u(x,y,0) & \text{for } (x,y)\in\left[-\mu_1 l\Delta t-\frac{\Delta x}{2},1-\mu_1 l\Delta t-\frac{\Delta x}{2}\right) \\
 & \times\left[-\mu_2 l\Delta t-\frac{\Delta y}{2},1-\mu_2 l\Delta t-\frac{\Delta y}{2}\right) \\
 & \setminus\bigcup_{j=1}^{|\hat{X}_l|}\left[\hat{x}_j-\frac{\Delta x}{2},\hat{x}_j+\frac{\Delta x}{2}\right)\times\left[\hat{y}_j-\frac{\Delta y}{2},\hat{y}_j+\frac{\Delta y}{2}\right),
\end{cases} \tag{5.57}
$$

and $v_l(x+1,y+1) = v_l(x,y)$ for all $(x,y) \in \mathbb{R}\times\mathbb{R}$, so is 1-periodic in both the x- and y-directions. The function $v_l(x,y)$ is equal to the function $u(x-\mu_1 l\Delta t, y-\mu_2 l\Delta t, 0)$ except within the rectangle $\left[\hat{x}_j - \frac{\Delta x}{2}, \hat{x}_j + \frac{\Delta x}{2}\right) \times \left[\hat{y}_j - \frac{\Delta y}{2}, \hat{y}_j + \frac{\Delta y}{2}\right)$ where the point $(\hat{x}_j, \hat{y}_j) \in \hat{X}_l$. The function constructs a rectangle based pyramid over each rectangle, with the apex of the pyramid taking the value of $u(\hat{x}_j, \hat{y}_j, 0)$ and the sides extending to the function $u(x,y,0)$. This makes the function continuous and equal to the function $u(x-\mu_1 l\Delta t, y-\mu_2 l\Delta t, 0)$ at every sample point of the domain. As $\hat{X}_l = \hat{X}_{[l]_{\mathrm{lcm}(b_1,b_2)}}$, $v_l(x,y) = v_{[l]_{\mathrm{lcm}(b_1,b_2)}}(x,y)$, so there are at most $\mathrm{lcm}(b_1,b_2)$ different functions $v_l(x,y)$ for $l = 0,\ldots,\mathrm{lcm}(b_1,b_2)-1$.

Now define a Fourier series for $v_l(x,y)$ for $l = 0,\ldots,\mathrm{lcm}(b_1,b_2)-1$, such that $\hat{X}_l \neq \emptyset$,

$$v_l(x,y) = \sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} d_{j,k}^{(l)} e^{2\pi i j x} e^{2\pi i k y}, \tag{5.58}$$

where $d_{j,k}^{(l)} \in \mathbb{C}$ for all $j,k \in \mathbb{Z}$. This is a convergent Fourier series for $v_l(x,y)$ by the conditions of the Lemma.

Let $c := \mathrm{lcm}(b_1,b_2)$ and apply the 2D DFT to $\tilde{\mathbf{y}}_l$. As for Lemma 3.12, when $\hat{X}_l = \emptyset$, the Fourier series of $u_0(x,y)$ should be considered, as seen in the calculations below. The same calculations are carried out using the Fourier series for $v_l(x,y)$ when $\hat{X}_l \neq \emptyset$. We are able to do this as by showing that $v_l(x,y) = v_{[l]_c}(x,y)$, we have shown that $d_{j,k}^{(l)} = d_{j,k}^{([l]_c)}$ for all $j,k \in \mathbb{Z}$.

$$\begin{aligned}
& \mathcal{F}_{p,q}(\tilde{\mathbf{y}}_l) \\
= \;& \sqrt{N_x N_y} \sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} c_{p-1+jN_x, q-1+kN_y} e^{-2\pi i(p-1+jN_x)\mu_1 l\Delta t} e^{-2\pi i(q-1+kN_y)\mu_2 l\Delta t}, \\
& \text{by the 2D Poisson summation,} \\
= \;& e^{\frac{-2\pi i(p-1)\mathrm{sgn}(\mu_1)h_1(l-[l]_c)}{N_x}} e^{\frac{-2\pi i(q-1)\mathrm{sgn}(\mu_2)h_2(l-[l]_c)}{N_y}} \mathcal{F}_{p,q}(\tilde{\mathbf{y}}_{[l]_c}), \qquad (5.59) \\
& \text{as } b_1 | (l-[l]_c) \text{ and } b_2 | (l-[l]_c),
\end{aligned}$$

We note that as,

$$\begin{aligned}
& e^{\frac{-2\pi i(p-1)\mathrm{sgn}(\mu_1)h_1(l-[l]_c)}{N_x}} e^{\frac{-2\pi i(q-1)\mathrm{sgn}(\mu_2)h_2(l-[l]_c)}{N_y}} \\
= \;& e^{\frac{-2\pi i(p-1+s_1N_x)\mathrm{sgn}(\mu_1)h_1(l-[l]_c)}{N_x}} e^{\frac{-2\pi i(q-1+s_2N_y)\mathrm{sgn}(\mu_2)h_2(l-[l]_c)}{N_y}},
\end{aligned}$$

for any $s_1, s_2 \in \mathbb{Z}$, we can rewrite (5.59) as,

$$\begin{aligned}
\mathcal{F}_{p,q}(\tilde{\mathbf{y}}_l) &= \begin{cases}
e^{\frac{-2\pi i(p-1)\mathrm{sgn}(\mu_1)h_1(l-[l]_c)}{N_x}} e^{\frac{-2\pi i(q-1)\mathrm{sgn}(\mu_2)h_2(l-[l]_c)}{N_y}} \mathcal{F}_{p,q}(\tilde{\mathbf{y}}_{[l]_c}), \\
\text{for } p = 1, \ldots, \frac{N_x+1}{2} \text{ and } q = 1, \ldots, \frac{N_y+1}{2}, \\
e^{\frac{-2\pi i(p-1)\mathrm{sgn}(\mu_1)h_1(l-[l]_c)}{N_x}} e^{\frac{2\pi i(N_y-q+1)\mathrm{sgn}(\mu_2)h_2(l-[l]_c)}{N_y}} \mathcal{F}_{p,q}(\tilde{\mathbf{y}}_{[l]_c}), \\
\text{for } p = 1, \ldots, \frac{N_x+1}{2} \text{ and } q = \frac{N_y+3}{2}, \ldots, N_y, \\
e^{\frac{2\pi i(N_x-p+1)\mathrm{sgn}(\mu_1)h_1(l-[l]_c)}{N_x}} e^{\frac{-2\pi i(q-1)\mathrm{sgn}(\mu_2)h_2(l-[l]_c)}{N_y}} \mathcal{F}_{p,q}(\tilde{\mathbf{y}}_{[l]_c}), \\
\text{for } p = \frac{N_x+3}{2}, \ldots, N_x \text{ and } q = 1, \ldots, \frac{N_y+1}{2}, \\
e^{\frac{2\pi i(N_x-p+1)\mathrm{sgn}(\mu_1)h_1(l-[l]_c)}{N_x}} e^{\frac{2\pi i(N_y-q+1)\mathrm{sgn}(\mu_2)h_2(l-[l]_c)}{N_y}} \mathcal{F}_{p,q}(\tilde{\mathbf{y}}_{[l]_c}), \\
\text{for } p = \frac{N_x+3}{2}, \ldots, N_x \text{ and } q = \frac{N_y+3}{2}, \ldots, N_y,
\end{cases} \\
&= \tilde{\lambda}_{p,q}^{l-[l]_c} \mathcal{F}_{p,q}(\tilde{\mathbf{y}}_{[l]_c}), \text{ for } p = 1, \ldots, N_x \text{ and } q = 1, \ldots, N_y.
\end{aligned} \tag{5.60}$$

We also have that,

$$\tilde{\lambda}_{p,q}^{l} \mathcal{F}_{p,q}(\tilde{\mathbf{x}}_0) = \tilde{\lambda}_{p,q}^{l-[l]_c} \tilde{\lambda}_{p,q}^{[l]_c} \mathcal{F}_{p,q}(\tilde{\mathbf{x}}_0), \text{ for } p = 1, \ldots, N_x \text{ and } q = 1, \ldots, N_y. \tag{5.61}$$

Therefore, substituting (5.60) and (5.61) into (5.51),

$$\begin{aligned}
& \mathcal{F}_{p,q}(\mathbf{r}_l) = \tilde{\lambda}_{p,q}^{l-[l]_c} \mathcal{F}_{p,q}(\mathbf{r}_{[l]_c}), \\
\Rightarrow \quad & V^* \mathbf{r}_l = \tilde{\Lambda}^{l-[l]_c} V^* \mathbf{r}_{[l]_c}, \\
\Rightarrow \quad & \mathbf{r}_l = \tilde{M}^{l-[l]_c} \mathbf{r}_{[l]_c}.
\end{aligned} \tag{5.62}$$

As $\tilde{\mathbf{y}}_0 = \tilde{\mathbf{x}}_0$, $\mathbf{r}_0 = \mathbf{0}$. Then by (5.62), when $[l]_c = 0$, $\mathbf{r}_l = \mathbf{0}$. Hence the result in (5.50). □

We can see from Lemma 5.6 that the MNIMC scheme for the 2D linear advection problem in (5.1), has a shifted $b_1\Delta t$-periodic nature in the x-direction and a shifted $b_2\Delta t$-periodic nature in the y-direction. This results in the scheme having a shifted $\mathrm{lcm}(b_1, b_2)\Delta t$-periodic nature. We can see this in Figure 5.2 where $h_1 = h_2 = \frac{1}{2}$. This gives that $b_1 = b_2 = 2$, so $\mathrm{lcm}(2, 2) = 2$ and we see that the 2D square function is recovered when the scheme is applied twice. If we now consider an example where $h_1 \neq h_2$ by setting $h_1 = \frac{1}{9}$ and $h_2 = \frac{1}{3}$, then we have that $\mathrm{lcm}(9, 3) = 9$. We would expect to see a shifted $9\Delta t$-periodic nature in the x-direction, a shifted $3\Delta t$-periodic nature in the y-direction and a shifted $9\Delta t$-periodic nature overall in the numerical results of the scheme. Examining Figure 5.5 we can see that this is the case.

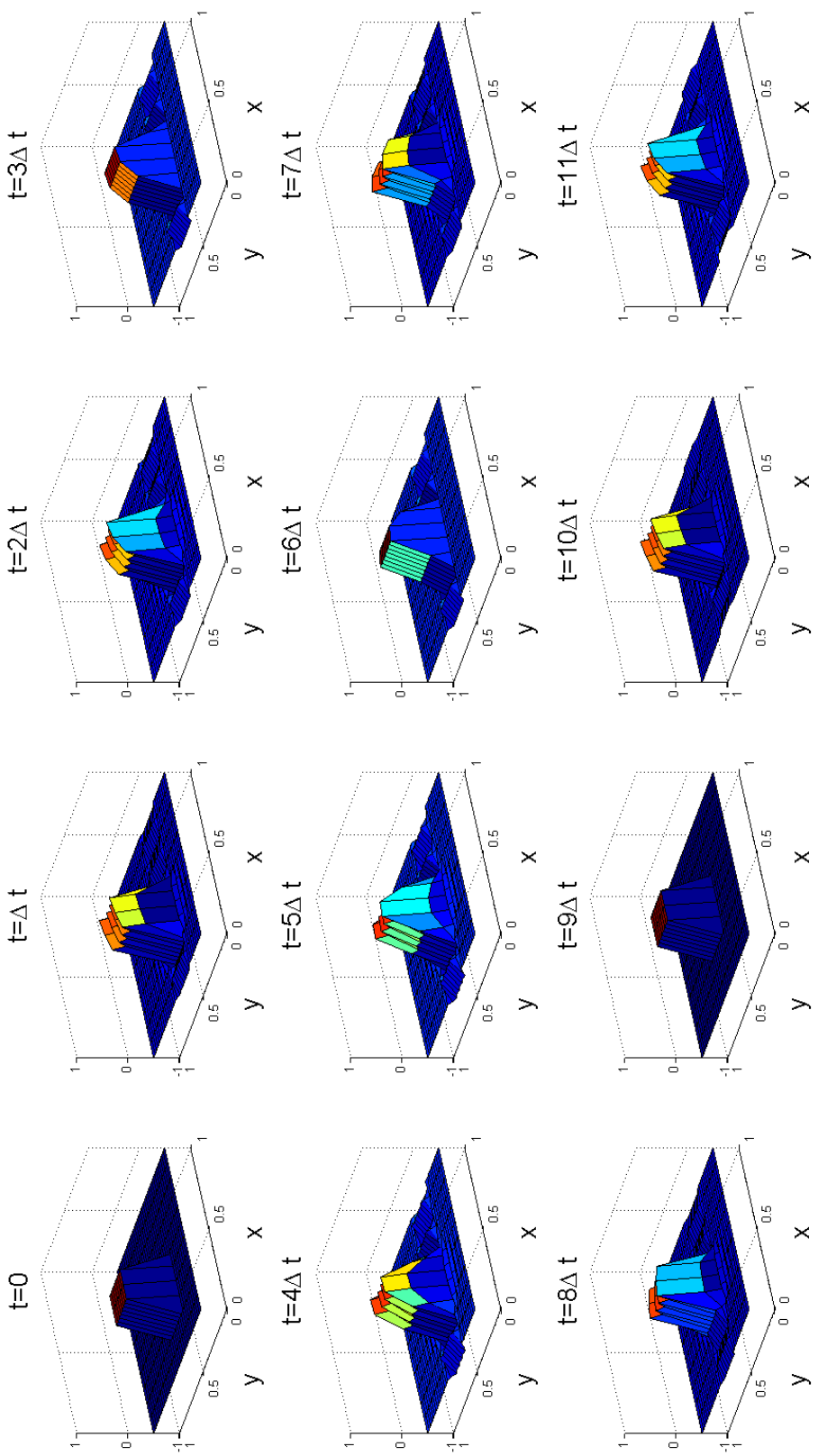


Figure 5.5: *The numerical results from applying the MNIMC scheme, for the 2D linear advection problem, to the 2D square function initial condition in (5.134). The aliasing errors in the scheme have a shifted $9\Delta t$-periodic nature in the x-direction, a shifted $3\Delta t$-periodic nature in the y-direction and an overall shifted $9\Delta t$-periodic nature. These results were generated using $N_x = 11$, $N_y = 33$, $\mu_1 = \mu_2 = 1$, $h_1 = \frac{1}{9}$ and $h_2 = \frac{1}{3}$ $\left(\Delta t = \frac{1}{99}\right)$.*

In Section 3.9 we found that the MNIMC scheme for the 1D linear advection equation, had the property that if the eigenvalues corresponding to a conjugate pair of the 1D DFT eigenvectors are swapped, the shifted periodic nature of the scheme is preserved. Suppose we also try this for the MNIMC scheme for the 2D linear advection problem, so that $(\overline{\tilde{\lambda}}_{p,q}, \mathbf{v}_{p,q})$ and $(\tilde{\lambda}_{p,q}, \overline{\mathbf{v}}_{p,q})$ form eigenpairs of the scheme, for some $p = 2, \ldots, N_x$ and $q = 2, \ldots, N_y$. We then find that this scheme is consistent, numerically stable and convergent and retains the shifted $\text{lcm}(b_1, b_2)\Delta t$-periodic nature.

It is possible to construct $2^{\frac{N_x N_y - 1}{2}}$ different schemes in this way for solving the 2D linear advection problem, that maintain the shifted $\text{lcm}(b_1, b_2)$-periodic property. These schemes would all be numerically non-dissipative with respect to all wavenumber components of the numerical solution for any values of h_1 and h_2, but would be numerically dispersive with respect to the resolvable wavenumber components where the eigenvalues have been swapped. The only configuration that is numerically non-dispersive with respect to all the resolvable wavenumber components, is that of the MNIMC scheme configuration.

Suppose we extend the 2D linear advection problem, to d-dimensions in space and construct an MNIMC scheme for this problem. Then based on our analysis on the 1D and 2D problems. we would expect the aliasing error in this scheme for the d-dimensional problem, to have a shifted $\text{lcm}(b_1, \ldots, b_d)\Delta t$-periodic nature. Here b_j is the denominator of $h_j \in \mathbb{Q}^+$, where $h_j := \frac{|\mu_j|\Delta t}{\Delta z_j}$. The variable μ_j is the constant wave speed of the d-dimensional problem in the z_j-direction, for $j = 1, \ldots, d$.

Now we have completed our analysis of the aliasing error introduced by the MNIMC scheme, we can use the scheme to construct perfect observations of the 2D linear advection problem. Therefore in the next Section, we begin our analysis of the strong constraint 4D-Var data assimilation problem set out in Section 2.3, where the 2D linear advection problem is our physical system of interest.

5.9 The effect of numerical dissipation and dispersion on the analysis vector

Throughout this chapter, we have seen that we can pose the numerical model error that enters into the solution of finite difference schemes for the 2D linear advection problem, in the same way as we did for the 1D linear advection problem. As a result, analysing the effects of numerical dissipation and dispersion on the analysis vector for the 2D problem is very similar to that of the 1D problem. As before in Section 3.10, we can construct the cost function for the 2D problem. Under the assumptions of Section 5.3, $\mathcal{M}_{l+1,l} := M$ and $\mathbf{x}_l \equiv \mathbf{U}^l$ for all l, and the cost function becomes,

$$J(\mathbf{x}_0) = \frac{1}{\sigma_o^2} \sum_{l=0}^{L} \left[\mathbf{y}_l - M^l \mathbf{x}_0\right]^T \left[\mathbf{y}_l - M^l \mathbf{x}_0\right]. \tag{5.63}$$

Our aim is to minimise (5.63) with respect to $\mathbf{x}_0$ and recover the discrete sample of $u_0(x, y)$, found in the vector $\tilde{\mathbf{x}}_0$. The formulation of this cost function is identical to that found in (3.67). As a result, the formulation of the analysis vector which minimises (5.63) with respect to $\mathbf{x}_0$ is identical to (3.68). Let $\mathbf{x}_a \in \mathbb{R}^{N_x N_y}$, denote the solution to our strong constraint 4D-Var data assimilation problem, ie: $\nabla J(\mathbf{x}_a) = 0$. Then,

$$\mathbf{x}_a = \left[\sum_{k=0}^{L}(M^T M)^k\right]^{-1}\sum_{l=0}^{L}(M^T)^l\mathbf{y}_l = V\left[\sum_{k=0}^{L}(\Lambda^*\Lambda)^k\right]^{-1}\sum_{l=0}^{L}(\Lambda^*)^l V^*\mathbf{y}_l, \tag{5.64}$$

using (5.15) and we would like to recover $\mathbf{x}_a = \tilde{\mathbf{x}}_0$. This can be re-written using the 2D DFT,

$$\mathcal{F}(\mathbf{x}_a) = \left[\sum_{k=0}^{L}(\Lambda^*\Lambda)^k\right]^{-1}\left[\sum_{l=0}^{L}(\Lambda^*)^l\mathcal{F}(\mathbf{y}_l)\right] = \left[I_N + \sum_{k=1}^{L}(\Lambda^*\Lambda)^k\right]^{-1}\left[\sum_{l=0}^{L}(\Lambda^*)^l\mathcal{F}(\mathbf{y}_l)\right]. \tag{5.65}$$

Equation (5.65) forms the coefficients for the 2D DFT basis in the construction of the analysis vector. Initially we wish to investigate the affects of numerical model error on the analysis vector in the absence of observation errors. Therefore, as discussed in Section 2.3, $\sigma_o^2 = 1$ is chosen in Equation (5.63). However, this does not affect the formulation of (5.64), using $\mathbf{y}_l = \tilde{\mathbf{y}}_l$.

Similarly to Lemma 3.13 for the 1D linear advection problem, the following Lemma expresses the analysis vector in terms of the sum of a matrix operation on $\tilde{\mathbf{x}}_0$, $A_L\tilde{\mathbf{x}}_0$ and an aliasing correction term $\boldsymbol{\rho}_L \in \mathbb{R}^{N_x N_y}$. The matrix $A_L \in \mathbb{R}^{N_x N_y \times N_x N_y}$ is constructed solely from the matrix M implementing the considered finite difference scheme for the problem and the matrix $\tilde{M}$ implementing the MNIMC scheme for the 2D linear advection problem. Lemma 5.7 is identical to Lemma 3.13 but with b replaced by $c \in \mathbb{N}$.

Lemma 5.7. *Let the assumptions of Lemma 5.6 hold true, allowing* $\mathbf{x}_a$ *to be stated as in (5.64). Consider perfect observations of the physical system ie:* $\mathbf{y}_l = \tilde{\mathbf{y}}_l$ *for all* $l = 0, \ldots, L$ *where* $L \in \mathbb{N}_0$ *is finite, in the form of (5.49). Then the analysis vector can be expressed as,*

$$\mathbf{x}_a = A_L\tilde{\mathbf{x}}_0 + \boldsymbol{\rho}_L, \tag{5.66}$$

where the model resolution matrix $A_L \in \mathbb{R}^{N_x N_y \times N_x N_y}$ *is such that,*

$$A_L = V\left[\sum_{k=0}^{L}(\Lambda^*\Lambda)^k\right]^{-1}\left[\sum_{l=0}^{L}(\Lambda^*\tilde{\Lambda})^l\right]V^*, \tag{5.67}$$

and $\boldsymbol{\rho}_L \in \mathbb{R}^{N_x N_y}$ *is given by,*

$$\begin{aligned}\boldsymbol{\rho}_L &= V\left[\sum_{k=0}^{L}(\Lambda^*\Lambda)^k\right]^{-1}\left[\left\{\sum_{l=0}^{\frac{L-[L]_c}{c}-1}(\Lambda^*\tilde{\Lambda})^{lc}\right\}\left\{\sum_{y=1}^{c-1}(\Lambda^*)^y V^*\mathbf{r}_y\right\}\right. \\ &\quad \left.+\left(\Lambda^*\tilde{\Lambda}\right)^{L-[L]_c}\left\{\sum_{y=1}^{[L]_c}(\Lambda^*)^y V^*\mathbf{r}_y\right\}\right], \end{aligned} \tag{5.68}$$

where $c := lcm(b_1, b_2)$. *Here we consider* $\sum_{j=1}^{0}(\Lambda^*)^j V^*\mathbf{r}_j = \mathbf{0}$ *and* $\sum_{l=0}^{-1}\left(\Lambda^*\tilde{\Lambda}\right)^{lc} = 0_{N_xN_y} \in \mathbb{R}^{N_xN_y \times N_xN_y}$ *as we assume* $\mathbf{r}_0 = \mathbf{0}$.

Proof. The proof is identical to that of Lemma 3.13, but makes use of the shifted $\text{lcm}(b_1, b_2)\Delta t$-periodic nature of the MNIMC scheme for the 2D linear advection problem, as detailed in Lemma 5.6. □

The eigenvalues of A_L in (5.67) determine the magnitude and phase change applied to each wavenumber component of $\tilde{\mathbf{x}}_0$ in the construction of $\mathbf{x}_a$. So as with the 1D problem, we can refer to them as *amplification factors* for the wavenumber components of $\tilde{\mathbf{x}}_0$. Let $\nu_{p,q}$ be an eigenvalue of A_L, $p = 1, \ldots, N_x$ and $q = 1, \ldots, N_y$. Due to the diagonal structures of Λ and $\tilde{\Lambda}$, $\nu_{p,q}$ is constructed solely from $\lambda_{p,q}$ and $\tilde{\lambda}_{p,q}$, the (p, q)th eigenvalues of M and $\tilde{M}$ respectively,

$$\nu_{p,q} = \frac{\sum_{l=0}^{L}\overline{\lambda}_{p,q}^{l}\tilde{\lambda}_{p,q}^{l}}{\sum_{k=0}^{L}|\lambda_{p,q}|^{2k}}. \tag{5.69}$$

Numerical model error enters into $\nu_{p,q}$ via both $\lambda_{p,q}$ and $\tilde{\lambda}_{p,q}$. In the case of $\tilde{\lambda}_{p,q}$ introduces numerical model error through the effects of aliasing. The complex conjugate properties of the eigenvalues of M and $\tilde{M}$, result in the same complex conjugate structure for the eigenvalues of A_L.

The quality of the analysis vector and the ability of the schemes to propagate unresolvable wavenumber components, can be explored similarly to Section 3.10. We would expect to see similar results to those for the 1D linear advection problem, such as the analysis vector for the numerically non-dissipative and dispersive Crank Nicolson scheme with respect to the resolvable wavenumber components of the numerical solution, losing information due to destructive interference between observations. However since the results of Chapter 5 yielded interesting results regarding a possible optimal number of discretisation points when considering full sets of observations to perform strong constraint 4D-Var data assimilation when considering full sets of observations, we perform a similar analysis in the remainder of this Chapter, to identify if a similar result extends to the 2D linear advection problem.

5.10 The spectral approach in the absence of observation errors

In Chapter 4, we explored the behaviour of the error in the analysis vector for the 1D linear advection equation, by constructing a bound for the l_2-norm of the error in the analysis vector. The bound was constructed using a spectral approach possibly allowing the dependence of the error on the smoothness of the true initial condition, the number of discretisation points when considering full sets of observations, the number of sets of observations in the assimilation window and the numerically dissipative and dispersive properties of the considered finite difference scheme, to be examined explicitly. It was found to be suitable for characterising the l_2-norm of the error in the analysis vector, with respect to the number of discretisation points when considering full sets of observations, for numerically dissipative and/or dispersive finite difference schemes with respect to the resolvable wavenumber components of the numerical solution. This revealed that if the true initial condition contains discontinuities, then the worst case is that the error does not decay as the number of discretisation points is increased. It is important to understand how considering a two-dimensional problem affects results like these and examine the behaviour of the error in the analysis vector for the 2D linear advection problem behaves. To this end, the same approach as in Chapter 4 will be taken to see if a similar bound can be constructed to characterise the behaviour of the l_2-norm of the error in the analysis vector, for the 2D linear advection problem.

The spectral method for constructing a bound for the l_2-norm of the error in the analysis vector for the 1D linear advection problem, required the use of a bound on the 1D Fourier coefficients and a bound on the error between the coefficient identified through the 1D DFT and the coefficient of the 1D Fourier series, for the same resolvable wavenumber component. As we wish to proceed in the same manner for the 2D problem, we require similar bounds, but in two-dimensions. However to the best of our knowledge, these bounds have not been extended to the 2D case. Therefore, we proceed to prove these theorems in the following sections.

5.10.1 A bound on the 2D Fourier coefficients

In the following Lemma, we prove a bound for the 2D Fourier coefficients, following the steps set out by Carslaw for the 1D case [61].

Lemma 5.8. *Let the function $f(x,y)$ be multiplicatively separable such that $f(x,y) = f_1(x)f_2(y)$, with $f_1(x)$ and $f_2(y)$ satisfying the conditions of Lemma 4.2 over $(0,T_1)$ with regularity $r_1 \in \mathbb{N}_0$ and $(0,T_2)$ with regularity $r_2 \in \mathbb{N}_0$ respectively. Define $Q_1, Q_2, Q_3, Q_4 \in \mathbb{N}$ as the number of rectangular sub-domains of $[0,T_1] \times [0,T_2]$, where $f_1(x)f_2(y)$, $f_1(x)f_2^{(r_2)}(y)$, $f_1^{(r_1)}(x)f_2(y)$ and $f_1^{(r_1)}(x)f_2^{(r_2)}(y)$ are continuous respectively, when $r_1 =$*

0 *and/or* $r_2 = 0$. *When* $r_1, r_2 \in \mathbb{N}$, *define* $Q_j = 1$ *for all* $j = 1, \ldots, 4$.

Then the coefficients of the Fourier series for $f(x, y)$, *given by* $f_{p,q}$, $p, q \in \mathbb{Z}$, *can be bounded such that,*

$$|f_{p,q}| \leq \begin{cases} A_1, & \text{for } p = q = 0, \\ \frac{A_2}{|q|^{r_2+1}}, & \text{for } p = 0 \text{ and } q \in \mathbb{Z} \setminus \{0\}, \\ \frac{A_3}{|p|^{r_1+1}}, & \text{for } p \in \mathbb{Z} \setminus \{0\} \text{ and } q = 0, \\ \frac{A_4}{|p|^{r_1+1}|q|^{r_2+1}} & \text{for } p, q \in \mathbb{Z} \setminus \{0\}. \end{cases} \tag{5.70}$$

where $A_1 = v_1 v_3$, $A_2 = \frac{4v_1 v_4 s_2 T_2^{r_2} Q_2}{(2\pi)^{r_2+1}}$, $A_3 = \frac{4v_2 v_3 s_1 T_1^{r_1} Q_3}{(2\pi)^{r_1+1}}$ *and* $A_4 = \frac{16 v_2 v_4 s_1 s_2 T_1^{r_1} T_2^{r_2} Q_4}{(2\pi)^{r_1+r_2+2}}$. *The variables* $v_1, v_2, v_3, v_4 \in \mathbb{R}$ *are the bounds on the functions* $f_1(x)$, $f_1^{(r_1)}(x)$, $f_2(y)$ *and* $f_2^{(r_2)}(y)$ *respectively and* $s_1, s_2 \in \mathbb{R}$ *are the number of monotone pieces* $f_1^{(r_1)}(x)$ *over* $(0, T_1)$ *and* $f_2^{(r_2)}(y)$ *over* $(0, T_2)$, *can be broken up into respectively.*

Proof. A corollary of Fubini's theorem for double integrals on rectangular domains is found in [84, p. 1010]. Given,

$$f_{p,q} = \frac{1}{T_1 T_2} \int_0^{T_1} \int_0^{T_2} f(x, y) e^{\frac{-2\pi i p x}{T_1}} e^{\frac{-2\pi i q y}{T_2}} \mathrm{d}y \, \mathrm{d}x, \tag{5.71}$$

we are integrating over a rectangular domain and $f(x, y)$ is multiplicatively separable into $f_1(x) e^{\frac{-2\pi i p x}{T_1}}$ and $f_2(y) e^{\frac{-2\pi i q y}{T_2}}$. If $r_1, r_2 \in \mathbb{N}$ then both functions are continuous on the rectangular domain, so the corollary allows (5.71) to be re-written as,

$$f_{p,q} = \frac{1}{T_1 T_2} \left(\int_0^{T_1} f_1(x) e^{\frac{-2\pi i p x}{T_1}} \mathrm{d}x \right) \left(\int_0^{T_2} f_2(y) e^{\frac{-2\pi i q y}{T_2}} \mathrm{d}y \right). \tag{5.72}$$

Applying the steps of the proof of Lemma A.4 in Appendix A, to each integral and bounding the result, we obtain (5.70) for $r_1, r_2 \in \mathbb{N}$ with $Q_j = 1$ for all $j = 1, \ldots, 4$.

Now consider the case of $r_1 = 0$ and $r_2 \in \mathbb{N}$. In this instance, $f_1(x)$ is not continuous on $[0, T_1]$, but $f_2(y)$ is continuous on $[0, T_2]$. Therefore, the integrand of (5.71) is not continuous on the rectangular domain $[0, T_1] \times [0, T_2]$ and we cannot apply the corollary of Fubini's theorem for double integrals to (5.71). Instead, we begin by applying integration by parts to the integral over y to obtain,

$$f_{p,q} = \begin{cases} \frac{1}{T_1 T_2} \int_0^{T_1} \int_0^{T_2} f_1(x) f_2(y) e^{\frac{-2\pi i p x}{T_1}} e^{\frac{-2\pi i q y}{T_2}} \mathrm{d}y \, \mathrm{d}x, \\ \text{for } p \in \mathbb{Z} \text{ and } q = 0, \\ \\ \frac{1}{T_1 T_2} \left(\frac{-i T_2}{2\pi q} \right)^{r_2} \int_0^{T_1} \int_0^{T_2} f_1(x) f_2^{(r_2)}(y) e^{\frac{-2\pi i p x}{T_1}} e^{\frac{-2\pi i q y}{T_2}} \mathrm{d}y \, \mathrm{d}x, \\ \text{for } p \in \mathbb{Z} \text{ and } q \in \mathbb{Z} \setminus \{0\}. \end{cases} \tag{5.73}$$

A multiplicatively separable function over a rectangular domain, composed of two piecewise continuous functions, is a piecewise continuous function whose continuous

pieces all have a rectangular domain. Therefore, we are able to partition $[0, T_1] \times [0, T_2]$ into Q_1 rectangular domains where $f_1(x)$ and $f_2(y)$ are both continuous and Q_2 rectangular domains where $f_1(x)$ and $f_2^{(r_2)}(y)$ are both continuous functions.

Define the partitions of $[0, T_1]\times[0, T_2]$ for $f_1(x)f_2(y)$ by $[(\hat{a}_1)_z, (\hat{b}_1)_z]\times[(\hat{c}_1)_z, (\hat{d}_1)_z] \subset [0, T_1] \times [0, T_2]$ for $z = 1, \ldots, Q_1$ and for $f_1(x)f_2^{(r_2)}(y)$ by $[(\hat{a}_2)_z, (\hat{b}_2)_z] \times [(\hat{c}_2)_z, (\hat{d}_2)_z] \subset [0, T_1] \times [0, T_2]$ for $z = 1, \ldots, Q_2$ such that,

$$\bigcup_{z=1}^{Q_j} [(\hat{a}_j)_z, (\hat{b}_j)_z] \times [(\hat{c}_j)_z, (\hat{d}_j)_z] = [0, T_1] \times [0, T_2],$$

for $j = 1, 2$. We are now able to apply the same corollary of Fubini's Theorem [84, p. 1010] to the double integral on each rectangular sub-domain,

$$f_{p,q} = \begin{cases} \frac{1}{T_1 T_2} \sum_{z=1}^{Q_1} \left(\int_{(\hat{a}_1)_z}^{(\hat{b}_1)_z} f_1(x) e^{\frac{-2\pi i p x}{T_1}} \mathrm{d}x \right) \left(\int_{(\hat{c}_1)_z}^{(\hat{d}_1)_z} f_2^{(r_1)}(y) e^{\frac{-2\pi i q y}{T_2}} \mathrm{d}y \right), \\ \text{for } p \in \mathbb{Z} \text{ and } q = 0, \\ \\ \frac{1}{T_1 T_2} \left(\frac{-iT_2}{2\pi q} \right)^{r_2} \sum_{z=1}^{Q_2} \left(\int_{(\hat{a}_2)_z}^{(\hat{b}_2)_z} f_1(x) e^{\frac{-2\pi i p x}{T_1}} \mathrm{d}x \right) \left(\int_{(\hat{c}_2)_z}^{(\hat{d}_2)_z} f_2^{(r_2)}(y) e^{\frac{-2\pi i q y}{T_2}} \mathrm{d}y \right), \\ \text{for } p \in \mathbb{Z} \text{ and } q \in \mathbb{Z} \setminus \{0\}. \end{cases} \tag{5.74}$$

Applying the steps of the proof of Lemma A.4 in Appendix A to each integral and bounding the result, we obtain,

$$|f_{p,q}| \leq \begin{cases} \frac{v_1 v_3}{T_1 T_2} \sum_{z=1}^{Q_1} |(\hat{b}_1)_z - (\hat{a}_1)_z||(\hat{d}_1)_z - (\hat{c}_1)_z|, & p = q = 0, \\ \frac{4 v_1 v_4 T_2^{r_2}}{T_1 (2\pi |q|)^{r_2+1}} \sum_{z=1}^{Q_2} |(\hat{b}_2)_z - (\hat{a}_2)_z|(s_2)_z, & p = 0 \text{ and } q \in \mathbb{Z} \setminus \{0\}, \\ \frac{4 v_2 v_3 T_1^{r_1}}{T_2 (2\pi)^{r_1+1} |p|^{r_1+1}} \sum_{z=1}^{Q_1} (s_1)_z |(\hat{d}_1)_z - (\hat{c}_1)_z|, & p \in \mathbb{Z} \setminus \{0\} \text{ and } q = 0, \\ \frac{16 v_2 v_4 T_1^{r_1} T_2^{r_2}}{(2\pi)^{r_1+r_2+2} |p|^{r_1+1} |q|^{r_2+1}} \sum_{z=1}^{Q_2} (s_3)_z (s_2)_z, & p, q \in \mathbb{Z} \setminus \{0\}. \end{cases} \tag{5.75}$$

As $r_1 = 0$, $Q_3 = Q_1$ and $Q_4 = Q_2$. Here $(s_1)_z \in \mathbb{N}$ is the number of monotone pieces of $f_1(x)$ over $[(\hat{a}_1)_z, (\hat{b}_1)_z]$. The variables $(s_2)_z, (s_3)_z \in \mathbb{N}$ are the number of monotone pieces of $f_1(x)$ over $[(\hat{a}_2)_z, (\hat{b}_2)_z]$ and $f_2^{(r_2)}(y)$ over $[(\hat{c}_2)_z, (\hat{d}_2)_z]$, respectively. As,

$$\sum_{z=1}^{Q_1} |(\hat{b}_1)_z - (\hat{a}_1)_z||(\hat{d}_1)_z - (\hat{c}_1)_z| = T_1 T_2,$$

$(s_1)_z \leq s_1$, $(s_2)_z \leq s_2$, $(s_3)_z \leq s_1$, $|(\hat{b}_2)_z - (\hat{a}_2)_z| \leq T_1$ for all $z = 1, \ldots, Q_1$ and $|(\hat{d}_1)_z - (\hat{c}_1)_z| \leq T_2$ for all $z = 1, \ldots, Q_2$, we obtain (5.70) for $r_1 = 0$ and $r_2 \in \mathbb{N}$. A similar approach is taken for the cases $r_1 = r_2 = 0$ and $r_1 \in \mathbb{N}$, $r_2 = 0$.

□

The constants A_1, A_2, A_3 and A_4 are independent of p and q. However it should be noted that A_2, A_3 and A_4 are in some way dependent upon the regularity of $u_0(x,y)$ in x and y, ie: r_1 and r_2 respectively. When $r_1 = 0$, $v_1 = v_2$, $Q_1 = Q_3$ and $Q_2 = Q_4$. Also, when $r_2 = 0$, $v_3 = v_4$, $Q_1 = Q_2$ and $Q_3 = Q_4$. Consequently, when $r_1 = r_2 = 0$, $Q_j = Q_k$ for all $j,k = 1,\ldots,4$. As we are using the proof of Lemma 4.2, it is not appropriate to consider $r_1 \to \infty$ or $r_2 \to \infty$.

We do not consider a non-separable $f(x,y)$ in Lemma 5.8, due to the challenges associated with it. In order to see some of these challenges, consider a non-separable $f(x,y)$. Let this function satisfy Definition 3.8, by replacing any references to differentiability with partial differentiability, such that $f(x,y)$ has regularity one in both the x- and y-directions over $(0,T_1)$ and $(0,T_2)$ respectively, together with boundary conditions $f(0,y) = f(T_1,y)$ for all $y \in [0,T_2]$ and $f(x,0) = f(x,T_2)$ for all $x \in [0,T_1]$. Then applying 2D integration by parts we obtain,

$$f_{p,q} = \begin{cases} \frac{1}{T_1T_2}\int_0^{T_1}\int_0^{T_2} f(x,y)\mathrm{d}y\,\mathrm{d}x, & \text{for } p = q = 0, \\ \frac{1}{T_1T_2}\left(\frac{-iT_2}{2\pi q}\right)\int_0^{T_1}\int_0^{T_2} f_y(x,y)e^{\frac{-2\pi iqy}{T_2}}\mathrm{d}y\,\mathrm{d}x, & \text{for } p = 0 \text{ and } q \in \mathbb{Z}\setminus\{0\}, \\ \frac{1}{T_1T_2}\left(\frac{-iT_1}{2\pi p}\right)\int_0^{T_1}\int_0^{T_2} f_x(x,y)e^{\frac{-2\pi ipx}{T_1}}\mathrm{d}y\,\mathrm{d}x, & \text{for } p \in \mathbb{Z}\setminus\{0\} \text{ and } q = 0. \end{cases} \tag{5.76}$$

However, when p and q are both non-zero, we obtain an ambiguity in the result. We could choose to integrate with respect to x,

$$f_{p,q} = \frac{1}{T_1T_2}\left(\frac{-iT_1}{2\pi p}\right)\int_0^{T_1}\int_0^{T_2} f_x(x,y)e^{\frac{-2\pi ipx}{T_1}}e^{\frac{-2\pi iqy}{T_2}}\mathrm{d}y\,\mathrm{d}x, \tag{5.77}$$

or with respect to y,

$$f_{p,q} = \frac{1}{T_1T_2}\left(\frac{-iT_2}{2\pi q}\right)\int_0^{T_1}\int_0^{T_2} f_y(x,y)e^{\frac{-2\pi ipx}{T_1}}e^{\frac{-2\pi iqy}{T_2}}\mathrm{d}y\,\mathrm{d}x, \tag{5.78}$$

for $p,q \in \mathbb{Z}\setminus\{0\}$. If we chose to form a bound for $f_{p,q}$ using (5.77), then we will capture the behaviour of $|f_{p,q}|$ with respect to p. If we instead form a bound using (5.78), we see that we will capture the behaviour of $|f_{p,q}|$ with respect to q, but not p. This, together with the result of Lemma 5.8, indicates that both (5.77) and (5.78) are inadequate for determining the behaviour of $|f_{p,q}|$ with respect to both p and q. This method of forming a bound is dependent upon the route taken to partially differentiate the function $f(x,y)$. This is inherently a bad idea, as the order of differentiation affects the number of times you can partially differentiate the function in each direction.

This in turn leads to the issue of defining the regularity of the function $f(x,y)$. A natural choice would be to modify Definition 3.8, so that all differentiable properties are partially differentiable properties, as we did with our example. This would suit a separable $f(x,y)$. However, the proof for a bound on $f_{p,q}$ for $p,q \in \mathbb{Z}\setminus\{0\}$, should be used to determine the definition.

Suppose we are able to form a bound for a non-separable $f(x, y)$, using the 2D integration by parts route. Then we may be left with trying to bound something similar to,

$$f_{p,q} = \frac{1}{T_1 T_2} \left(\frac{-iT_1}{2\pi p} \right)^{r_1} \left(\frac{-iT_2}{2\pi q} \right)^{r_2} \int_0^{T_1} \int_0^{T_2} f_{r_1 x, r_2 y}(x, y) e^{\frac{-2\pi i p x}{T_1}} e^{\frac{-2\pi i q y}{T_2}} \mathrm{d}y \, \mathrm{d}x, \quad (5.79)$$

for $p, q \in \mathbb{Z} \setminus \{0\}$ and some $r_1, r_2 \in \mathbb{N}_0$. In the 1D case, once integration by parts had been applied as many times as possible, the *2nd Mean Value Theorem for integrals* in one-dimension was used. We would like to do the same in this problem. It is not possible to apply this theorem to each of our integrals individually. To demonstrate this, suppose we apply the theorem to the inner integral in (5.79). This results in the limits of the inner integrals over each monotone piece, becoming dependent on the independent variable of the outer integral, in this case x. The number of monotone pieces the inner integral is split into, is different for each x and the calculation becomes very complex as a result. The only piece of literature we were able to find regarding a possible 2D version of this proof, was that of Young [85]. We hope that this paper will aid in the proof of a bound on $f_{p,q}$, for $p, q \in \mathbb{Z} \setminus \{0\}$. We leave such a proof as future work.

5.10.2 A bound on the error in the 2D DFT

In this Section, we consider the error in the coefficient found by the 2D DFT, when compared with the Fourier coefficient of the 2D Fourier series for the same resolvable wavenumber component. The creation of this bound for the 1D DFT by Henson [66], required considering continuous and discontinuous functions separately, due to the use of the Poisson summation in the former. Here we will do similar, following the idea of Henson's proofs [66]. In Lemma 5.9 we consider $r_1, r_2 \in \mathbb{N}$ and in Lemma 5.10, we consider $r_1 = 0$ and/or $r_2 = 0$.

Lemma 5.9. *Let the assumptions of Lemma 5.8 hold true, replacing the reference to Lemma 4.2 with one to Lemma 4.3. Also let* $\mathbf{F} \in \mathbb{R}^{N_x N_y}$ *be such that* $\{\mathbf{F}\}_{(k-1)N_x+j} = f(x_{j-1}, y_{k-1})$ *where* $x_{j-1} = (j-1)\Delta x$ *and* $y_{k-1} = (k-1)\Delta y$ *for* $j = 1, \ldots, N_x$ *and*

$k = 1, \ldots, N_y$ *such that* $\Delta x = \frac{T_1}{N_x}$ *and* $\Delta y = \frac{T_2}{N_y}$. *Then,*

$$\left|\frac{1}{\sqrt{N_x N_y}}\mathcal{F}_{p,q}(\mathbf{F}) - f_{p-1,q-1}\right|$$

$$\leq \begin{cases} \frac{B_1}{N_x^{r_1+1}N_y^{r_2+1}} + \frac{B_2}{N_x^{r_1+1}} + \frac{B_3}{N_y^{r_2+1}}, & \text{for } p = q = 1, \\ \frac{B_1}{N_x^{r_1+1}N_y^{r_2+1}} + \frac{B_4}{N_x^{r_1+1}|q-1|^{r_2+1}} + \frac{B_3}{N_y^{r_2+1}}, & \text{for } p = 1 \\ & \text{and } q = 2, \ldots, \left\lfloor \frac{N_y}{2} \right\rfloor + 1, \\ \frac{B_1}{N_x^{r_1+1}N_y^{r_2+1}} + \frac{B_2}{N_x^{r_1+1}} + \frac{B_5}{|p-1|^{r_1+1}N_y^{r_2+1}}, & \text{for } p = 2, \ldots, \left\lfloor \frac{N_x}{2} \right\rfloor + 1 \\ & \text{and } q = 1, \\ \frac{B_1}{N_x^{r_1+1}N_y^{r_2+1}} + \frac{B_4}{N_x^{r_1+1}|q-1|^{r_2+1}} + \frac{B_5}{N_y^{r_2+1}|p-1|^{r_1+1}}, & \text{for } p = 2, \ldots, \left\lfloor \frac{N_x}{2} \right\rfloor + 1 \\ & \text{and } q = 2, \ldots, \left\lfloor \frac{N_y}{2} \right\rfloor + 1. \end{cases} \tag{5.80}$$

$$\left|\frac{1}{\sqrt{N_x N_y}}\mathcal{F}_{p,q}(\mathbf{F}) - f_{p-1,-N_y+q-1}\right|$$

$$\leq \begin{cases} \frac{B_1}{N_x^{r_1+1}N_y^{r_2+1}} + \frac{B_4}{N_x^{r_1+1}|q-N_y-1|^{r_2+1}} + \frac{B_3}{N_y^{r_2+1}}, & \text{for } p = 1 \\ & \text{and } q = \left\lfloor \frac{N_y}{2} \right\rfloor + 2, \ldots, N_y, \\ \frac{B_1}{N_x^{r_1+1}N_y^{r_2+1}} + \frac{B_4}{N_x^{r_1+1}|q-N_y-1|^{r_2+1}} + \frac{B_5}{|p-1|^{r_1+1}N_y^{r_2+1}}, & \text{for } p = 2, \ldots, \left\lfloor \frac{N_x}{2} \right\rfloor + 1 \\ & \text{and } q = \left\lfloor \frac{N_y}{2} \right\rfloor + 2, \ldots, N_y. \end{cases} \tag{5.81}$$

$$\left|\frac{1}{\sqrt{N_x N_y}}\mathcal{F}_{p,q}(\mathbf{F}) - f_{-N_x+p-1,q-1}\right|$$

$$\leq \begin{cases} \frac{B_1}{N_x^{r_1+1}N_y^{r_2+1}} + \frac{B_2}{N_x^{r_1+1}} + \frac{B_5}{|p-N_x-1|^{r_1+1}N_y^{r_2+1}}, & \text{for } p = \left\lfloor \frac{N_x}{2} \right\rfloor + 2, \ldots, N_x \\ & \text{and } q = 1, \\ \frac{B_1}{N_x^{r_1+1}N_y^{r_2+1}} + \frac{B_4}{N_x^{r_1+1}|q-1|^{r_2+1}} + \frac{B_5}{|p-N_x-1|^{r_1+1}N_y^{r_2+1}}, & \text{for } p = \left\lfloor \frac{N_x}{2} \right\rfloor + 2, \ldots, N_x \\ & \text{and } q = 2, \ldots, \left\lfloor \frac{N_y}{2} \right\rfloor + 1. \end{cases} \tag{5.82}$$

$$\left|\frac{1}{\sqrt{N_x N_y}}\mathcal{F}_{p,q}(\mathbf{F}) - f_{-N_x+p-1,-N_y+q-1}\right| \leq \frac{B_1}{N_x^{r_1+1}N_y^{r_2+1}} + \frac{B_4}{N_x^{r_1+1}|q-N_y-1|^{r_2+1}} + \frac{B_5}{|p-N_x-1|^{r_1+1}N_y^{r_2+1}}, \tag{5.83}$$

for $p = \left\lfloor \frac{N_x}{2} \right\rfloor + 2, \ldots, N_x$ *and* $q = \left\lfloor \frac{N_y}{2} \right\rfloor + 2, \ldots, N_y$, *where* $B_1 = A_4 C(r_1+1)C(r_2+1)$,

$B_2 = A_3C(r_1+1)$, $B_3 = A_2C(r_2+1)$, $B_4 = A_4C(r_1+1)$ *and* $B_5 = A_4C(r_2+1)$ *and* $C : \mathbb{N} \to \mathbb{R}$ *such that* $r \mapsto 2^r + \zeta(r)$.

Proof. Consider the following 2D continuous Fourier series for $f(x, y)$,

$$f(x,y) = \sum_{p=-\infty}^{\infty} \sum_{q=-\infty}^{\infty} f_{p,q} e^{\frac{2\pi i p x}{T_1}} e^{\frac{2\pi i q y}{T_2}}, \tag{5.84}$$

where,

$$f_{p,q} = \frac{1}{T_1 T_2} \int_0^{T_1} \int_0^{T_2} f(x,y) e^{\frac{-2\pi i p x}{T_1}} e^{\frac{-2\pi i q y}{T_2}} \,\mathrm{d}y \,\mathrm{d}x. \tag{5.85}$$

Under the conditions of the Lemma, the Fourier series for $f(x, y)$ is convergent. As $r_1, r_2 \in \mathbb{N}$, $f(x, y)$ is continuous at every point in the domain so is equal to its Fourier series. The 2D Poisson summation in Section 5.3.2 can then be used to express the 2D DFT of $f(x, y)$,

$$\frac{1}{\sqrt{N_x N_y}} \mathcal{F}_{p,q}(\mathbf{F}) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} f_{jN_x+p-1, kN_y+q-1}. \tag{5.86}$$

We wish to find a bound on the error between (5.86) and the coefficient for the corresponding resolvable wavenumber component of the Fourier series of $f(x, y)$. In order to achieve this we require the following bounds determined by Henson [66] in his proof for the error in the 1D DFT coefficients for $r_1, r_2 \in \mathbb{N}$,

$$\sum_{j=1}^{\infty} \left(\frac{1}{|jN_x + p - 1|^{r_1+1}} + \frac{1}{|-jN_x + p - 1|^{r_1+1}} \right) \leq \frac{C(r_1+1)}{N_x^{r_1+1}}, \tag{5.87}$$

$$\sum_{k=1}^{\infty} \left(\frac{1}{|kN_y + q - 1|^{r_2+1}} + \frac{1}{|-kN_y + q - 1|^{r_2+1}} \right) \leq \frac{C(r_2+1)}{N_y^{r_2+1}}. \tag{5.88}$$

- Initially, consider $p = 1, \ldots, \left\lfloor \frac{N_x}{2} \right\rfloor + 1$ and $q = 1, \ldots, \left\lfloor \frac{N_y}{2} \right\rfloor + 1$. The corresponding Fourier series coefficient is $f_{p-1,q-1}$. Then we bound,

$$\left| \frac{1}{\sqrt{N_x N_y}} \mathcal{F}_{p,q}(\mathbf{F}) - f_{p-1,q-1} \right| = \left| \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} f_{jN_x+p-1, kN_y+q-1} - f_{p-1,q-1} \right|. \tag{5.89}$$

  Following the steps of Henson's proof for the error in the 1D DFT coefficient and applying the bound developed in Lemma 5.8, together with (5.87) and (5.88), we obtain (5.80).

- Next consider the case of $p = 1, \ldots, \left\lfloor \frac{N_x}{2} \right\rfloor + 1$ and $q = \left\lfloor \frac{N_y}{2} \right\rfloor + 2, \ldots, N_y$. The corresponding Fourier series coefficient is $f_{p-1,-N_y+q-1}$. Then applying the same analysis as for the previous case, we obtain (5.81).

- The next case is to consider $p = \left\lfloor \frac{N_x}{2} \right\rfloor + 2, \ldots, N_x$ and $q = 1, \ldots, \left\lfloor \frac{N_y}{2} \right\rfloor + 1$.

The corresponding Fourier coefficient is $f_{-N_x+p-1,q-1}$. Then applying the same analysis as for the previous case, we obtain (5.82).

- The final case is to consider $p = \lfloor \frac{N_x}{2} \rfloor + 2, \ldots, N_x$ and $q = \left\lfloor \frac{N_y}{2} \right\rfloor + 2, \ldots, N_y$. The corresponding Fourier series coefficient is $f_{-N_x+p-1,-N_y+q-1}$. Then applying the same analysis as for the previous case, we obtain (5.83).

□

The result of Lemma 5.9 can be written more succinctly by using the notation of Henson [66]. Lemma 5.9 identifies a bound for the error in the 2D DFT for a multiplicatively separable function $f(x,y) = f_1(x)f_2(y)$, where the regularities of $f_1(x)$ and $f_2(y)$ are such that $r_1, r_2 \in \mathbb{N}$. We require a similar bound for this function when $r_1 = 0$ and/or $r_2 = 0$. This will be derived in Lemma 5.10.

Lemma 5.10. *Let the assumptions of Lemma 5.9 hold true, but let $f_1(x)$ over $(0,T_1)$ and $f_2(y)$ over $(0,T_2)$ have regularities r_1 and r_2 respectively, with $r_1, r_2 \in \mathbb{N}_0$ such that at least one of them is equal to zero. Then Equations (5.80)-(5.83) of Lemma 5.9 hold true for $r_1 = 0$ and/or $r_2 = 0$, with,*

$$B_1 = \begin{cases} A_4C(r_1+1)C(r_2+1), & \text{for } r_1 = 0 \text{ and } r_2 \in \mathbb{N} \text{ or } r_1 \in \mathbb{N} \text{ and } r_2 = 0, \\ A_4C(r_1+1)C(r_2+1) - A_9, & \text{for } r_1 = r_2 = 0, \end{cases} \tag{5.90}$$

$$B_2 = \begin{cases} A_3C(r_1+1), & \text{for } r_1 \in \mathbb{N} \text{ and } r_2 = 0, \\ A_3C(r_1+1) + A_5, & \text{for } r_1 = 0 \text{ and } r_2 \in \mathbb{N}_0, \end{cases} \tag{5.91}$$

$$B_3 = \begin{cases} A_2C(r_2+1), & \text{for } r_1 = 0 \text{ and } r_2 \in \mathbb{N}, \\ A_2C(r_2+1) + A_6, & \text{for } r_1 \in \mathbb{N}_0 \text{ and } r_2 = 0, \end{cases} \tag{5.92}$$

$$B_4 = \begin{cases} A_4C(r_1+1), & \text{for } r_1 \in \mathbb{N} \text{ and } r_2 = 0, \\ A_4C(r_1+1) + A_7, & \text{for } r_1 = 0,\ r_2 \in \mathbb{N}_0, \end{cases} \tag{5.93}$$

$$B_5 = \begin{cases} A_4C(r_2+1), & \text{for } r_1 = 0 \text{ and } r_2 \in \mathbb{N}, \\ A_4C(r_2+1) + A_8, & \text{for } r_1 \in \mathbb{N}_0 \text{ and } r_2 = 0, \end{cases} \tag{5.94}$$

where A_1, A_2, A_3 and A_4 are defined in Lemma 5.9 and $A_5 := \frac{2v_1v_3w_1}{T_1}$, $A_6 := \frac{2v_1v_3w_2}{T_2}$, $A_7 = \frac{4v_1v_4s_2T_2^{r_2}w_1}{T_1(2\pi)^{r_2+1}}$, $A_8 := \frac{4v_2v_3s_1T_1^{r_1}w_2}{T_2(2\pi)^{r_1+1}}$ and $A_9 := \frac{2v_1v_3w_1w_2}{T_1T_2}$.

Proof. In order to bound the error in the 2D DFT of $f(x,y) = f_1(x)f_2(y)$ when $r_1 = 0$ and/or $r_2 = 0$, we follow the ideas in the proof for the bound on the error in the 1D DFT when $r = 0$, as set out by Henson [66] and corrected in Section A.2 of Appendix A. When $r_1 = 0$ and/or $r_2 = 0$, $f(x,y)$ contains at least one discontinuity, therefore we cannot use the Poisson summation as in Lemma 5.9, to construct the 2D DFT of

$f(x,y)$ and hence bound the error in it. Instead we define a new function $g : \mathbb{R} \times \mathbb{R}$ such that $(x,y) \mapsto g(x,y)$, that possesses the same 2D DFT as $f(x,y)$, is continuous, has a convergent Fourier series and is multiplicatively separable.

Define $g(x,y) = g_1(x)g_2(y)$ where $g_1, g_2 : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ such that $x \mapsto g_1(x)$ and $y \mapsto g_2(y)$. If $r_1 = 0$, we define $g_1(x)$ by interpolating $f_1(x)$ and if $r_2 = 0$, we define $g_2(y)$ by interpolating $f_2(y)$. The interpolation technique is identical to that described in Henson [66] and corrected in Section A.2 of Appendix A. If $r_1 \in \mathbb{N}$, then $g_1(x) := f_1(x)$. Similarly, if $r_2 \in \mathbb{N}$, then $g_2(y) := f_2(y)$.

Now consider the error in the 2D DFT of the function $f(x,y)$, adding and subtracting the coefficients for the Fourier series of $g(x,y)$, given by $g_{p-1,q-1} \in \mathbb{C}$ for $p, q \in \mathbb{Z}$. For example, in the case of $p = 1, \ldots, \left\lfloor \frac{N_x}{2} \right\rfloor + 1$ and $q = 1, \ldots, \left\lfloor \frac{N_y}{2} \right\rfloor + 1$,

$$\begin{aligned} & \left| \frac{1}{\sqrt{N_x N_y}} \mathcal{F}_{p,q}(\mathbf{F}) - f_{p-1,q-1} \right|, \\ = & \left| \frac{1}{\sqrt{N_x N_y}} \mathcal{F}_{p,q}(\mathbf{F}) - g_{p-1,q-1} + g_{p-1,q-1} - f_{p-1,q-1} \right|, \\ \leq & \left| \frac{1}{\sqrt{N_x N_y}} \mathcal{F}_{p,q}(\mathbf{F}) - g_{p-1,q-1} \right| + \left| g_{p-1,q-1} - f_{p-1,q-1} \right|. \end{aligned} \tag{5.95}$$

The function $g(x,y)$ is multiplicatively separable with $r_1, r_2 \in \mathbb{N}$, has a convergent Fourier series and the same 2D DFT as $f(x,y)$, so Lemma 5.9 is then satisfied and can be used to bound,

$$\left| \frac{1}{\sqrt{N_x N_y}} \mathcal{F}_{p,q}(\mathbf{F}) - g_{p-1,q-1} \right|. \tag{5.96}$$

The same method is applied for the remaining values of p and q, using the corresponding resolvable Fourier coefficient to $\mathcal{F}_{p,q}(\mathbf{F})$.

Next we require a bound for $|g_{p,q} - f_{p,q}|$ when $p = -\left\lfloor \frac{N_x}{2} \right\rfloor, \ldots, \left\lfloor \frac{N_x}{2} \right\rfloor$ and $q = -\left\lfloor \frac{N_y}{2} \right\rfloor, \ldots, \left\lfloor \frac{N_y}{2} \right\rfloor$, as these are the corresponding resolvable wavenumber coefficients

of the Fourier series, to the 2D DFT. Consider the case of $r_1 = 0$ and $r_2 \in \mathbb{N}$. Then,

$$
\begin{aligned}
& |g_{p,q} - f_{p,q}|, \\
= \; & \frac{1}{T_1 T_2} \left| \int_0^{T_1} \int_0^{T_2} (g(x,y) - f(x,y))\, e^{\frac{-2\pi i p x}{T_1}} e^{\frac{-2\pi i q y}{T_2}} \,\mathrm{d}y \,\mathrm{d}x \right|, \\
= \; & \frac{1}{T_1 T_2} \left| \sum_{z=1}^{w_1} \int_{\hat{a}_z}^{\hat{b}_z} (g_1(x) - f_1(x))\, e^{\frac{-2\pi i p x}{T_1}} \int_0^{T_2} f_2(y) e^{\frac{-2\pi i q y}{T_2}} \,\mathrm{d}y \,\mathrm{d}x \right|, \\
& \text{as } g_2(y) := f_2(y) \text{ when } r_2 \in \mathbb{N}, \\
= \; & \begin{cases} \frac{1}{T_1 T_2} \left| \sum_{z=1}^{w_1} \int_{\hat{a}_z}^{\hat{b}_z} (g_1(x) - f_1(x))\, e^{\frac{-2\pi i p x}{T_1}} \int_0^{T_2} f_2(y) \mathrm{d}y \,\mathrm{d}x \right|, \\ \text{for } p = -\left\lfloor \frac{N_x}{2} \right\rfloor, \ldots, \left\lfloor \frac{N_x}{2} \right\rfloor \text{ and } q = 0, \\ \\ \frac{T_2^{r_2 - 1}}{T_1 (2\pi |q|)^{r_2}} \left| \sum_{z=1}^{w_1} \int_{\hat{a}_z}^{\hat{b}_z} (g_1(x) - f_1(x))\, e^{\frac{-2\pi i p x}{T_1}} \int_0^{T_2} f_2^{(r_2)}(y) e^{\frac{-2\pi i q y}{T_2}} \,\mathrm{d}y \,\mathrm{d}x \right|, \\ \text{for } p = -\left\lfloor \frac{N_x}{2} \right\rfloor, \ldots, \left\lfloor \frac{N_x}{2} \right\rfloor \text{ and } q = -\left\lfloor \frac{N_y}{2} \right\rfloor, \ldots, -1, 1, \ldots, \left\lfloor \frac{N_y}{2} \right\rfloor, \end{cases}
\end{aligned}
\tag{5.97}
$$

Here $w_1 \in \mathbb{N}$ is the number of sub-domains $[x_j, x_{j+1}]$ for $j = 0, \ldots, N_x - 1$, where $f_1(x)$ contains a discontinuity and $[\hat{a}_z, \hat{b}_z]$ denotes each of these subdomains for $z = 1, \ldots, w_1$.

As the integrals in (5.97) are defined on a rectangular domain $[\hat{a}_z, \hat{b}_z] \times [0, T_2]$ and the functions $f_1(x)$ and $f_2^{(r_2)}(y)$ are piecewise continuous functions on this domain, the rectangular domain can be partitioned into $P_1, P_2 \in \mathbb{N}$ rectangular domains where $f_1(x) f_2(y)$ and $f_1(x) f_2^{(r_2)}(y)$ are continuous respectively. Define the partition of $[\hat{a}_z, \hat{b}_z] \times [0, T_2]$ for $(g_1(x) - f_1(x)) f_2(y)$ by $[(\hat{\alpha}_1)_j, (\hat{\beta}_1)_j] \times [(\hat{\gamma}_1)_j, (\hat{\delta}_1)_j]$ for $j = 1, \ldots, P_1$ and for $(g_1(x) - f_1(x)) f_2^{(r_2)}(y)$ by $[(\hat{\alpha}_2)_j, (\hat{\beta}_2)_j] \times [(\hat{\gamma}_2)_j, (\hat{\delta}_2)_j]$ for $j = 1, \ldots, P_2$ such that,

$$
\bigcup_{j=1}^{P_k} [(\hat{\alpha}_k)_j, (\hat{\beta}_k)_j] \times [(\hat{\gamma}_k)_j, (\hat{\delta}_k)_j] = [\hat{a}_z, \hat{b}_z] \times [0, T_2],
$$

for $k = 1, 2$. The integrands of the integrals in (5.97) are then continuous and multiplicatively separable over these rectangular domains, so the Corollary from Fubini's Theorem for double integrals over rectangular domains [84, p. 1010] used in Lemma 5.8 can be applied. This results in,

$$
\begin{aligned}
& |g_{p,q} - f_{p,q}|, \\
= \; & \begin{cases} \frac{1}{T_1 T_2} \left| \sum_{z=1}^{w_1} \sum_{j=1}^{P_1} \left(\int_{(\hat{\alpha}_1)_j}^{(\hat{\beta}_1)_j} (g_1(x) - f_1(x))\, e^{\frac{-2\pi i p x}{T_1}} \mathrm{d}x \right) \left(\int_{(\hat{\gamma}_1)_j}^{(\hat{\delta}_1)_j} f_2(y) \mathrm{d}y \right) \right|, \\ \text{for } p = -\left\lfloor \frac{N_x}{2} \right\rfloor, \ldots, \left\lfloor \frac{N_x}{2} \right\rfloor \text{ and } q = 0, \\ \\ \frac{T_2^{r_2 - 1}}{T_1 (2\pi |q|)^{r_2}} \left| \sum_{z=1}^{w_1} \sum_{j=1}^{P_2} \left(\int_{(\hat{\alpha}_2)_j}^{(\hat{\beta}_2)_z} (g_1(x) - f_1(x))\, e^{\frac{-2\pi i p x}{T_1}} \mathrm{d}x \right) \left(\int_{(\hat{\gamma}_2)_j}^{(\hat{\delta}_2)_j} f_2^{(r_2)}(y) e^{\frac{-2\pi i q y}{T_2}} \mathrm{d}y \right) \right|, \\ \text{for } p = -\left\lfloor \frac{N_x}{2} \right\rfloor, \ldots, \left\lfloor \frac{N_x}{2} \right\rfloor \text{ and } q = -\left\lfloor \frac{N_y}{2} \right\rfloor, \ldots, -1, 1, \ldots, \left\lfloor \frac{N_y}{2} \right\rfloor, \end{cases}
\end{aligned}
\tag{5.98}
$$

Applying the *2nd Mean Value Theorem (MVT) for integrals* [86] stated in Lemma A.3

of Appendix A, to the integral over y and bounding the result similarly to Lemma 5.8 we obtain,

$$\begin{aligned} & |g_{p,q} - f_{p,q}|, \\ \leq & \begin{cases} \frac{2v_1 v_3 w_1 \Delta x}{T_1}, \\ \text{for } p = -\left\lfloor \frac{N_x}{2} \right\rfloor, \ldots, \left\lfloor \frac{N_x}{2} \right\rfloor \text{ and } q = 0, \\ \\ \frac{4v_1 v_4 s_2 T_2^{r_2} w_1 \Delta x}{T_1 (2\pi|q|)^{r_2+1}}, \\ \text{for } p = -\left\lfloor \frac{N_x}{2} \right\rfloor, \ldots, \left\lfloor \frac{N_x}{2} \right\rfloor \text{ and } q = -\left\lfloor \frac{N_y}{2} \right\rfloor, \ldots, -1, 1, \ldots, \left\lfloor \frac{N_y}{2} \right\rfloor, \end{cases} \\ & \text{by the generalised Fundamental Theorem of Calculus.} \end{aligned} \tag{5.99}$$

As $g_1(x)$ is a linear interpolation of $f_1(x)$, it is also bounded by v_1. A similar analysis for $r_1 \in \mathbb{N}$ and $r_2 = 0$ produces,

$$\begin{aligned} & |g_{p,q} - f_{p,q}|, \\ \leq & \begin{cases} \frac{2v_1 v_3 w_2 \Delta y}{T_2}, \\ \text{for } p = 0 \text{ and } q = -\left\lfloor \frac{N_y}{2} \right\rfloor, \ldots, \left\lfloor \frac{N_y}{2} \right\rfloor \\ \\ \frac{4v_2 v_3 s_1 T_1^{r_1} w_2 \Delta y}{T_2 (2\pi|p|)^{r_1+1}}, \\ \text{for } p = -\left\lfloor \frac{N_x}{2} \right\rfloor, \ldots, -1, 1, \ldots, \left\lfloor \frac{N_x}{2} \right\rfloor \text{ and } q = -\left\lfloor \frac{N_y}{2} \right\rfloor, \ldots, \left\lfloor \frac{N_y}{2} \right\rfloor, \end{cases} \end{aligned} \tag{5.100}$$

where $w_2 \in \mathbb{N}$ is the number of sub-domains $[y_k, y_{k+1}]$ for $k = 0, \ldots, N_y - 1$, where $f_2(y)$ contains a discontinuity. Let the domain $[\hat{c}_s, \hat{d}_s]$ denote these subdomains for $s = 1, \ldots, w_2$. Then when considering the case of $r_1 = r_2 = 0$,

$$
\begin{aligned}
& |g_{p,q} - f_{p,q}|, \\
= \; & \frac{1}{T_1 T_2} \left| \sum_{z=1}^{w_1} \int_{\hat{a}_z}^{\hat{b}_z} \int_0^{T_2} (g_1(x)g_2(y) - f_1(x)f_2(y))\, e^{\frac{-2\pi i p x}{T_1}} e^{\frac{-2\pi i q y}{T_2}} \mathrm{d}y\, \mathrm{d}x \right|, && (5.101) \\
& + \frac{1}{T_1 T_2} \left| \sum_{s=1}^{w_2} \int_0^{T_1} \int_{\hat{c}_s}^{\hat{d}_s} (g_1(x)g_2(y) - f_1(x)f_2(y))\, e^{\frac{-2\pi i p x}{T_1}} e^{\frac{-2\pi i q y}{T_2}} \mathrm{d}y\, \mathrm{d}x \right|, && (5.102) \\
& - \frac{1}{T_1 T_2} \left| \sum_{z=1}^{w_1} \sum_{s=1}^{w_2} \int_{\hat{a}_z}^{\hat{b}_z} \int_{\hat{c}_s}^{\hat{d}_s} (g_1(x)g_2(y) - f_1(x)f_2(y))\, e^{\frac{-2\pi i p x}{T_1}} e^{\frac{-2\pi i q y}{T_2}} \mathrm{d}y\, \mathrm{d}x \right|, && (5.103) \\
\leq \; & \begin{cases}
\frac{2v_1 v_3 w_1 \Delta x}{T_1} + \frac{2v_1 v_3 w_2 \Delta y}{T_2} - \frac{2v_1 v_3 \Delta x \Delta y w_1 w_2}{T_1 T_2}, \\
\text{for } p = q = 0, \\
\\
\frac{4v_1 v_4 s_2 T_2^{r_2} w_1 \Delta x}{T_1 (2\pi|q|)^{r_2+1}} + \frac{2v_1 v_3 w_2 \Delta y}{T_2} - \frac{2v_1 v_3 \Delta x \Delta y w_1 w_2}{T_1 T_2}, \\
\text{for } p = 0 \text{ and } q = -\left\lfloor \frac{N_y}{2} \right\rfloor, \ldots, -1, 1, \ldots, \left\lfloor \frac{N_y}{2} \right\rfloor, \\
\\
\frac{2v_1 v_3 w_1 \Delta x}{T_1} + \frac{4v_2 v_3 s_1 T_1^{r_1} w_2 \Delta y}{T_2 (2\pi|p|)^{r_1+1}} - \frac{2v_1 v_3 \Delta x \Delta y w_1 w_2}{T_1 T_2}, \\
\text{for } p = -\left\lfloor \frac{N_x}{2} \right\rfloor, \ldots, -1, 1, \ldots, \left\lfloor \frac{N_x}{2} \right\rfloor \text{ and } q = 0, \\
\\
\frac{4v_1 v_4 s_2 T_2^{r_2} w_1 \Delta x}{T_1 (2\pi|q|)^{r_2+1}} + \frac{4v_2 v_3 s_1 T_1^{r_1} w_2 \Delta y}{T_2 (2\pi|p|)^{r_1+1}} - \frac{2v_1 v_3 \Delta x \Delta y w_1 w_2}{T_1 T_2}, \\
\text{for } p = -\left\lfloor \frac{N_x}{2} \right\rfloor, \ldots, -1, 1, \ldots, \left\lfloor \frac{N_x}{2} \right\rfloor \text{ and } q = -\left\lfloor \frac{N_y}{2} \right\rfloor, \ldots, -1, 1, \ldots, \left\lfloor \frac{N_y}{2} \right\rfloor.
\end{cases} && (5.104)
\end{aligned}
$$

The 2nd MVT for integrals was not applied to the integral in (5.103), so the generalised FTC was applied and the integral bounded directly. Then by combining the result of Lemma 5.9 with (5.95), (5.99), (5.100) and (5.104), we obtain the result of this Lemma. □

Suppose the function $f(x, y)$ is a multiplicatively non-separable function and contains a discontinuity. Constructing a similar proof to Lemma 5.10, we construct a new function $g(x, y)$ defined using $f(x, y)$ as a starting point, interpolating across the discontinuities in $f(x, y)$. The corresponding Fourier coefficient of $g(x, y)$ is then added and subtracted as in (5.95), allowing the equivalent of Lemma 5.9 for non-separable functions, to be applied. A simple formulation for $g(x, y)$ is constructed by introducing square-based pyramids into $f(x, y)$. Given a point of discontinuity, identify the four grid squares within $\frac{\Delta x}{2}$ and $\frac{\Delta y}{2}$ of the discontinuity. Then using the mid-point of each of these squares evaluated in $f(x, y)$ as the four corners of the base of the pyramid and $f(x, y)$ at the intersection of these four grid squares as the apex, construct the square-based pyramid. This creates a continuous $g(x, y)$ which evaluates identically to $f(x, y)$ at each grid point.

The problem with this formulation arises when $f(x, y)$ is discontinuous in only one direction. Suppose a line of discontinuity exists parallel to the y-direction for some $x_0 \in [0, T_1)$. Then $f(x, y)$ is continuous for each x over $\mathbb{R}$ and discontinuous in the y-direction at $x = x_0$. Given the described properties of $f(x, y)$, even though regularity has not been defined for a multiplicatively non-separable function, it is imaginable that the regularity in the x-direction would be $r_1 \in \mathbb{N}$ and in the y-direction would be $r_2 = 0$. We desire that $g(x, y)$ have non-zero regularities in both directions. Defining $g(x, y)$ by interpolating $f(x, y)$ over the grid squares containing discontinuities in $[0, T_1] \times [0, T_2]$, as discussed above, results in $g(x, y)$ being continuous in the x- and y-directions. Therefore we would expect that $r_1, r_2 \in \mathbb{N}$. However $g(x, y)$ may not have the same regularity in the x-direction as $f(x, y)$. This is not consistent with the results of Lemma 5.10, so the choice of $g(x, y)$ may need to be reconsidered for such functions.

The constants B_1, B_2, B_3, B_4 and B_5 of Lemmas 5.9 and 5.10 are independent of p, q, N_x and N_y. The coefficient B_1 is dependent on r_1 and r_2, whilst the coefficients B_2 and B_4 are dependent on r_1 and independent of r_2 and the coefficients B_3 and B_5 are independent of r_1, but dependent on r_2. The results of these Lemmas cannot be considered as either $r_1 \to \infty$ or $r_2 \to \infty$ as they depend on Lemma 5.8 which cannot be considered under these conditions.

Now we have the bounds on the 2D continuous Fourier coefficients and the error in the 2D DFT when compared to its corresponding 2D Fourier coefficient, we can follow the proof of Lemma 4.4 to construct a bound for the error in the analysis vector.

5.10.3 A bound on the error in the analysis vector

Now we have the bounds on the 2D Fourier coefficients and the error in the 2D DFT for problem (5.1). Using these, we can take the same spectral approach to find a bound on the error in the analysis vector, as found in Lemma 4.5, for the 1D problem in (3.1).

Lemma 5.11. *Let the assumptions of Lemma 5.10 hold for the function $u_0(x, y)$ defined in problem (5.1), over its domain $[0, 1) \times [0, 1)$ and those detailed so far allowing $\mathbf{x}_a$ to be stated as in (5.66), hold true. Also, let $\tilde{\mathbf{x}}_l$ be defined as in Section 5.6.1 for all*

$l = 0, \ldots, L$. *Then,*

$$\begin{aligned}
&\|\tilde{\mathbf{x}}_0 - \mathbf{x}_a\|_2^2 \\
\leq\ & N_x N_y \left\{ |1-\nu_{1,1}| A_1 + (|1-\nu_{1,1}| + 2\xi_{1,1}) \left(\frac{B_1}{N_x^{r_1+1} N_y^{r_2+1}} + \frac{B_2}{N_x^{r_1+1}} + \frac{B_3}{N_y^{r_2+1}} \right) \right\}^2 \\
&+ 2N_x N_y \sum_{q=2}^{\frac{N_y+1}{2}} \left\{ |1-\nu_{1,q}| \frac{A_2}{|q-1|^{r_2+1}} + (|1-\nu_{1,q}| + 2\xi_{1,q}) \left(\frac{B_1}{N_x^{r_1+1} N_y^{r_2+1}} \right.\right. \\
&\left.\left. + \frac{B_4}{N_x^{r_1+1}|q-1|^{r_2+1}} + \frac{B_3}{N_y^{r_2+1}} \right) \right\}^2 \\
&+ 2N_x N_y \sum_{p=2}^{\frac{N_x+1}{2}} \left\{ |1-\nu_{p,1}| \frac{A_3}{|p-1|^{r_1+1}} + (|1-\nu_{p,1}| + 2\xi_{p,1}) \left(\frac{B_1}{N_x^{r_1+1} N_y^{r_2+1}} \right.\right. \\
&\left.\left. + \frac{B_2}{N_x^{r_1+1}} + \frac{B_5}{|p-1|^{r_1+1} N_y^{r_2+1}} \right) \right\}^2 \\
&+ 4N_x N_y \sum_{p=2}^{\frac{N_x+1}{2}} \sum_{q=2}^{\frac{N_y+1}{2}} \left\{ |1-\nu_{p,q}| \frac{A_4}{|p-1|^{r_1+1}|q-1|^{r_2+1}} \right. \\
&\left. + (|1-\nu_{p,q}| + 2\xi_{p,q}) \left(\frac{B_1}{N_x^{r_1+1} N_y^{r_2+1}} + \frac{B_4}{N_x^{r_1+1}|q-1|^{r_2+1}} + \frac{B_5}{|p-1|^{r_1+1} N_y^{r_2+1}} \right) \right\}^2
\end{aligned} \tag{5.105}$$

where B_1, B_2, B_3, B_4 *and* B_5 *are defined as in Lemmas 5.9 and 5.10. We also define* $\xi_{p,q} \in \mathbb{R}$ *for* $p = 1, \ldots, N_x$ *and* $q = 1, \ldots, N_y$ *by,*

$$\xi_{p,q} = \frac{\left| \sum_{l=0}^{\frac{L-[L]_c}{c}-1} \left[|\lambda_{p,q}|^c e^{ic\phi_{p,q}} \right]^l \right| \left\{ \sum_{y=1}^{c-1} |\lambda_{p,q}|^y \right\} + |\lambda_{p,q}|^{L-[L]_c} \sum_{y=1}^{[L]_c} |\lambda_{p,q}|^y}{\sum_{k=0}^{L} |\lambda_{p,q}|^{2k}}, \tag{5.106}$$

where $c = lcm(b_1, b_2)$.

Proof. Following the outline of Lemma 4.4, using Equation (5.66) of Lemma 5.7 results in,

$$\|\tilde{\mathbf{x}}_0 - \mathbf{x}_a\|_2^2 = \|(I - A_L)\tilde{\mathbf{x}}_0 - \boldsymbol{\rho}_L\|_2^2 \leq \sum_{p=1}^{N_x} \sum_{q=1}^{N_y} \left\{ |1-\nu_{p,q}| \, |\mathcal{F}_{p,q}(\tilde{\mathbf{x}}_0)| + |\mathcal{F}_{p,q}(\boldsymbol{\rho}_L)| \right\}^2. \tag{5.107}$$

We now require a bound for $|\mathcal{F}_{p,q}(\tilde{\mathbf{x}}_0)|$ and $|\mathcal{F}_{p,q}(\boldsymbol{\rho}_L)|$ for each (p, q). Using the same

method as in Lemma 4.4,

$$
\begin{aligned}
&|\mathcal{F}_{p,q}(\tilde{\mathbf{x}}_0)| \\
&\leq \begin{cases}
\sqrt{N_xN_y}\left\{A_1 + \frac{B_1}{N_x^{r_1+1}N_y^{r_2+1}} + \frac{B_2}{N_x^{r_1+1}} + \frac{B_2}{N_y^{r_2+1}}\right\} & \text{for } p=1 \text{ and } q=1 \\
\sqrt{N_xN_y}\left\{\frac{A_2}{|q-1|^{r_2+1}} + \frac{B_1}{N_x^{r_1+1}N_y^{r_2+1}} + \frac{B_4}{N_x^{r_1+1}|q-1|^{r_2+1}} + \frac{B_3}{N_y^{r_2+1}}\right\} & \text{for } p=1 \text{ and } q=2,\ldots,\frac{N_y+1}{2} \\
\sqrt{N_xN_y}\left\{\frac{A_3}{|p-1|^{r_1+1}} + \frac{B_1}{N_x^{r_1+1}N_y^{r_2+1}} + \frac{B_2}{N_x^{r_1+1}} + \frac{B_5}{|p-1|^{r_1+1}N_y^{r_2+1}}\right\} & \text{for } p=2,\ldots,\frac{N_x+1}{2} \text{ and } q=1 \\
\sqrt{N_xN_y}\left\{\frac{A_4}{|p-1|^{r_1+1}|q-1|^{r_2+1}} + \frac{B_1}{N_x^{r_1+1}N_y^{r_2+1}} + \frac{B_4}{N_x^{r_1+1}|q-1|^{r_2+1}} + \frac{B_5}{|p-1|^{r_1+1}N_y^{r_2+1}}\right\} & \text{for } p=2,\ldots,\frac{N_x+1}{2} \text{ and } q=2,\ldots,\frac{N_y+1}{2} \\
\sqrt{N_xN_y}\left\{\frac{A_2}{|q-1-N_y|^{r_2+1}} + \frac{B_1}{N_x^{r_1+1}N_y^{r_2+1}} + \frac{B_4}{N_x^{r_1+1}|q-N_y-1|^{r_2+1}} + \frac{B_3}{N_y^{r_2+1}}\right\} & \text{for } p=1 \text{ and } q=\frac{N_y+3}{2},\ldots,N_y \\
\sqrt{N_xN_y}\left\{\frac{A_4}{|p-1|^{r_1+1}|q-1-N_y|^{r_2+1}} + \frac{B_1}{N_x^{r_1+1}N_y^{r_2+1}} + \frac{B_4}{N_x^{r_1+1}|q-N_y-1|^{r_2+1}} + \frac{B_5}{|p-1|^{r_1+1}N_y^{r_2+1}}\right\} & \text{for } p=2,\ldots,\frac{N_x+1}{2} \text{ and } q=\frac{N_y+3}{2},\ldots,N_y \\
\sqrt{N_xN_y}\left\{\frac{A_3}{|p-1-N_x|^{r_1+1}} + \frac{B_1}{N_x^{r_1+1}N_y^{r_2+1}} + \frac{B_2}{N_x^{r_1+1}} + \frac{B_5}{|p-N_x-1|^{r_1+1}N_y^{r_2+1}}\right\} & \text{for } p=\frac{N_x+3}{2},\ldots,N_x \text{ and } q=1 \\
\sqrt{N_xN_y}\left\{\frac{A_4}{|p-1-N_x|^{r_1+1}|q-1|^{r_2+1}} + \frac{B_1}{N_x^{r_1+1}N_y^{r_2+1}} + \frac{B_4}{N_x^{r_1+1}|q-1|^{r_2+1}} + \frac{B_5}{|p-N_x-1|^{r_1+1}N_y^{r_2+1}}\right\} & \text{for } p=\frac{N_x+3}{2},\ldots,N_x \text{ and } q=2,\ldots,\frac{N_y+1}{2} \\
\sqrt{N_xN_y}\left\{\frac{A_4}{|p-1-N_x|^{r_1+1}|q-1-N_y|^{r_2+1}} + \frac{B_1}{N_x^{r_1+1}N_y^{r_2+1}} + \frac{B_4}{N_x^{r_1+1}|q-N_y-1|^{r_2+1}} + \frac{B_5}{|p-N_x-1|^{r_1+1}N_y^{r_2+1}}\right\} & \text{for } p=\frac{N_x+3}{2},\ldots,N_x \text{ and } q=\frac{N_y+3}{2},\ldots,N_y
\end{cases}
\end{aligned}
\tag{5.108}
$$

Now consider $|\mathcal{F}_{p,q}(\boldsymbol{\rho}_L)|$ where $\boldsymbol{\rho}_L$ is defined in (5.68). As in Lemma 4.5, we apply the triangle inequality to create a bound in terms of $|\mathcal{F}_{p,q}(\mathbf{r}_l)|$ for $l = 0, \dots [l]_{\mathrm{lcm}(b_1,b_2)}$ for finite L,

$$\begin{aligned} & |\mathcal{F}_{p,q}(\boldsymbol{\rho}_L)| \\ \leq \; & \frac{\left|\sum_{l=0}^{\frac{L-[L]_c}{c}-1} \left(\overline{\lambda}_{p,q}\tilde{\lambda}_{p,q}\right)^{lc}\right| \left\{\sum_{j=1}^{c-1} |\lambda_{p,q}|^j |\mathcal{F}_{p,q}(\mathbf{r}_j)|\right\} + |\lambda_{p,q}|^{L-[L]_c} \left\{\sum_{j=1}^{[L]_c} |\lambda_{p,q}|^j |\mathcal{F}_{p,q}(\mathbf{r}_j)|\right\}}{\sum_{r=0}^{L} |\lambda_p|^{2r}}, \end{aligned} \tag{5.109}$$

where $c = \mathrm{lcm}(b_1, b_2)$. Now consider $|\mathcal{F}_{p,q}(\mathbf{r}_j)|$ for $j = 1, \dots, \mathrm{lcm}(b_1, b_2)$. As in Lemma 4.5, we use the result of Lemmas 5.9 and 5.10,

$$
\begin{aligned}
& |\mathcal{F}_{p,q}(\mathbf{r}_j)| \\
= \; & \left|\mathcal{F}_{p,q}(\tilde{\mathbf{y}}_j) - \tilde{\lambda}_{p,q}^{j}\mathcal{F}_{p,q}(\tilde{\mathbf{x}}_0)\right| \\
\leq \; & \begin{cases}
2\sqrt{N_xN_y}\left\{\frac{B_1}{N_x^{r_1+1}N_y^{r_2+1}} + \frac{B_2}{N_x^{r_1+1}} + \frac{B_3}{N_y^{r_2+1}}\right\}, & \text{for } p=q=1, \\
2\sqrt{N_xN_y}\left\{\frac{B_1}{N_x^{r_1+1}N_y^{r_2+1}} + \frac{B_4}{N_x^{r_1+1}|q-1|^{r_2+1}} + \frac{B_5}{N_y^{r_2+1}}\right\}, & \text{for } p=1 \text{ and } q=2,\ldots,\frac{N_y+1}{2}, \\
2\sqrt{N_xN_y}\left\{\frac{B_1}{N_x^{r_1+1}N_y^{r_2+1}} + \frac{B_2}{N_x^{r_1+1}} + \frac{B_5}{|p-1|^{r_1+1}N_y^{r_2+1}}\right\}, & \text{for } p=2,\ldots,\frac{N_x+1}{2} \text{ and } q=1, \\
2\sqrt{N_xN_y}\left\{\frac{B_1}{N_x^{r_1+1}N_y^{r_2+1}} + \frac{B_4}{N_x^{r_1+1}|q-1|^{r_2+1}} + \frac{B_5}{N_y^{r_2+1}|p-1|^{r_1+1}}\right\}, & \text{for } p=2,\ldots,\frac{N_x+1}{2} \text{ and } q=2,\ldots,\frac{N_y+1}{2}. \\
2\sqrt{N_xN_y}\left\{\frac{B_1}{N_x^{r_1+1}N_y^{r_2+1}} + \frac{B_4}{N_x^{r_1+1}|q-N_y-1|^{r_2+1}} + \frac{B_3}{N_y^{r_2+1}}\right\}, & \text{for } p=1 \text{ and } q=\frac{N_y+3}{2},\ldots,N_y, \\
2\sqrt{N_xN_y}\left\{\frac{B_1}{N_x^{r_1+1}N_y^{r_2+1}} + \frac{B_4}{N_x^{r_1+1}|q-N_y-1|^{r_2+1}} + \frac{B_5}{|p-1|^{r_1+1}N_y^{r_2+1}}\right\}, & \text{for } p=2,\ldots,\frac{N_x+1}{2} \text{ and } q=\frac{N_y+3}{2},\ldots,N_y. \\
2\sqrt{N_xN_y}\left\{\frac{B_1}{N_x^{r_1+1}N_y^{r_2+1}} + \frac{B_2}{N_x^{r_1+1}} + \frac{B_5}{|p-N_x-1|^{r_1+1}N_y^{r_2+1}}\right\}, & \text{for } p=\frac{N_x+3}{2},\ldots,N_x \text{ and } q=1, \\
2\sqrt{N_xN_y}\left\{\frac{B_1}{N_x^{r_1+1}N_y^{r_2+1}} + \frac{B_4}{N_x^{r_1+1}|q-1|^{r_2+1}} + \frac{B_5}{|p-N_x-1|^{r_1+1}N_y^{r_2+1}}\right\}, & \text{for } p=\frac{N_x+3}{2},\ldots,N_x \text{ and } q=2,\ldots,\frac{N_y+1}{2}. \\
2\sqrt{N_xN_y}\left\{\frac{B_1}{N_x^{r_1+1}N_y^{r_2+1}} + \frac{B_4}{N_x^{r_1+1}|q-N_y-1|^{r_2+1}} + \frac{B_5}{|p-N_x-1|^{r_1+1}N_y^{r_2+1}},\right\}, & \text{for } p=\frac{N_x+3}{2},\ldots,N_x \text{ and } q=\frac{N_y+3}{2},\ldots,N_y,
\end{cases}
\end{aligned}
\tag{5.110}
$$

Let $\xi_{p,q}$ be defined as (5.106) for $p = 1,\ldots,N_x$ and $q = 1,\ldots,N_y$. Then $|\mathcal{F}_{p,q}(\boldsymbol{\rho}_L)|$ is bounded by (5.110), multiplied by $\xi_{p,q}$. Due to the complex conjugate pair nature of $\lambda_{p,q}$ described in Section 5.3.1, $\xi_{p,q}$ has the same complex conjugate structure. Substituting the bound on $|\mathcal{F}_{p,q}(\boldsymbol{\rho}_L)|$ and (5.108) into (5.107) results in (5.105).

□

This Lemma provides a bound for the l_2-norm of the error in the analysis vector under the influence of errors introduced by finite difference approximations in the for-

ward model solving problem (5.1). Comparing the bounds in (5.105) and (4.19) for the 2D and 1D linear advection problems respectively, we can see similarities between their formulations. The bound is explicitly dependent on the regularity of the initial condition $u_0(x, y)$ in both the x- and y-directions (r_1 and r_2 respectively), the numerically dissipative and dispersive properties of the finite different scheme used as the forward model ($\nu_{p,q}$), the number of discretisation points in the x- and y-directions (N_x and N_y respectively) when considering full sets of observations and the number of sets of observations in the assimilation window (L). The similarities between the two bounds is not surprising as their proofs are produced using the same ideas. As a result, it is likely that the bound in Equation (5.105) can be interpreted as representing the worst case behaviour of the l_2-norm of the error in the analysis vector, similarly to the way the bound in Equation (4.19) was discussed in Section 4.3.7.

The next step in our analysis is to compare the behaviour of the bound against the results from strong constraint 4D-Var numerical experiments. This will allow us to determine if the bound is suitable for characterising the behaviour of the l_2-norm of the error in the analysis vector.

5.11 Analysis of the Bound

In this Section we perform a similar analysis to Section 4.3, to determine if the bound in Equation (5.105) is suitable for characterising the behaviour of the l_2-norm of the error in the analysis vector for the 2D linear advection problem. The behaviour of the l_2-norm of the error in the analysis vector can be identified through strong constraint 4D-Var data assimilation numerical experiments, whilst the behaviour of the bound can be identified through analysing the behaviour of its constituent summations both numerically and analytically. The behaviour of these quantities can be described through analysing their orders of convergence to zero with respect to N_x, N_y and L for varying regularity initial conditions.

In order to find these orders of convergence, we need to vary the values of these variables. However, whilst doing this, we need to maintain the numerically dissipative and dispersive properties of these schemes. As we have seen in the Tables in Section 5.5.3, these properties are determined by the values of h_1 and h_2. Suppose we fix h_1 and h_2 such that our numerical scheme is numerically stable, using some $\Delta x = \frac{1}{N_x}$, $\Delta y = \frac{1}{N_y}$ and Δt. Now suppose we decrease Δx by increasing N_x, $N_x^{(1)} = 3N_x$. Then in order to keep h_1 constant, we must decrease Δt, $\Delta t^{(1)} = \frac{\Delta t}{3}$. This decreases the length of the assimilation window given by $L\Delta t$ and also affects h_2. The knock on affect is then that N_y must be increased by the same factor as N_x to ensure that h_2 remains constant. Therefore when investigating the order of convergence of the error in the analysis vector with respect to N_x, we must also investigate the order of convergence with respect to N_y to preserve the numerically dissipative and dispersive properties of the schemes. This means that we are unable to investigate the order of convergence

with respect to N_x and N_y individually.

Section 5.3 investigated the numerically dissipative and dispersive properties of the Upwind and Crank-Nicolson schemes for solving problem (5.1). Since the Upwind scheme is numerically dissipative and dispersive with respect to resolvable wavenumber components and the Crank-Nicolson scheme is always numerically non-dissipative and dispersive with respect to resolvable wavenumber components, this does not lead us to obvious values to choose for h_1 and h_2, to perform our experiments. We require that $h_1 + h_2 = h \leq 1$, to maintain numerical stability of the Upwind scheme. The easiest way to investigate the order of convergence with respect to N_x and N_y is to set $N_x = N_y$ and to choose N_x to be odd so the MNIMC scheme can be used to define $|1 - \nu_{p,q}|$ terms for $p = 1, \ldots, N_x$ and $q = 1, \ldots, N_y$. The result of this is that $h_1 = h_2$. So we choose $h_1 = h_2 = \frac{1}{2}$, which results in $h = 1$. Then both schemes are numerically stable when increasing N_x and N_y. We also choose $\mu_1 = \mu_2 = 1$.

The number of observations is given by $N_x N_y L$, so increasing any of these factors increases the number of observations. Increasing N_x and N_y results in $\Delta t = \frac{h_1}{\mu_1 N_x} = \frac{h_2}{\mu_2 N_y}$ decreasing. As observations are taken at every point in space, every Δt, this results in the density of observations being increased in both space and time as N_x and N_y are increased.

The bound in Equation (5.105) can be re-written in the form of a sum of summations, similarly to the way the bound in 4.19 was re-written in the form of Equation (4.41). The Upwind and Crank-Nicolson schemes both have the property that $|1 - \nu_{1,1}| = 0$. Taking this into account, the bound can be re-written in the form of a sum comprised of more than 40 summations whose order of convergence to zero needs to be analysed with respect to N_x and L. In order to minimise the analysis involved, we use the results from the Section 4.3 to guide our choice in summations to be analysed. Section 4.3 showed that the summation,

$$S_1 = N_x \sum_{p=2}^{\frac{N_x+1}{2}} \frac{|1 - \nu_p|^2}{(p-1)^{2(r+1)}}$$

was the dominant summation of the equivalent bound on the l_2-norm of the error in the analysis vector for the 1D linear advection problem. We choose to analyse summations of a similar form in the construction of Equation 5.105 to see if their order of convergence both analytically and numerically, represents the numerical orders of convergence to zero for the l_2-norm of the error in the analysis vector, found through

strong constraint 4D-Var numerical experiments. These summations are,

$$R_1 := N_x N_y \sum_{p=2}^{\frac{N_x+1}{2}} \sum_{q=2}^{\frac{N_y+1}{2}} \frac{|1-\nu_{p,q}|^2}{|p-1|^{2(r_1+1)}|q-1|^{2(r_2+1)}}, \tag{5.111}$$

$$R_2 := N_x N_y \sum_{p=2}^{\frac{N_x+1}{2}} \frac{|1-\nu_{p,1}|^2}{|p-1|^{2(r_1+1)}}, \tag{5.112}$$

$$R_3 := N_x N_y \sum_{q=2}^{\frac{N_y+1}{2}} \frac{|1-\nu_{1,q}|^2}{|q-1|^{2(r_2+1)}}. \tag{5.113}$$

We now investigate the order of convergence of $|1-\nu_{p,q}|$ with respect to N_x and L.

5.11.1 The order of convergence of $|1-\nu_{p,q}|$

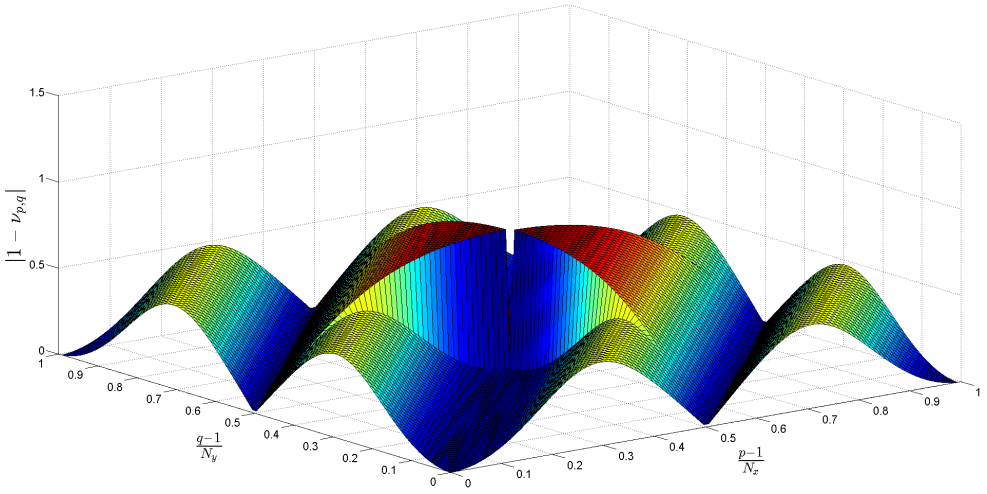
The $|1-\nu_{p,q}|$ terms arise in the bound in Equation (5.105) as a contribution from the considered finite difference scheme. These coefficients are dependent on N_x, N_y and L; N_x and N_y determine the number of points, whilst L determines the shape of the plots in Figure 5.6. For the reasons stated in Section 4.3.1, we consider L to be small in comparison to N_x and N_y whilst analysing the behaviour of $|1-\nu_{p,q}|$ with respect to N_x and L.

The order of convergence with respect to N_x

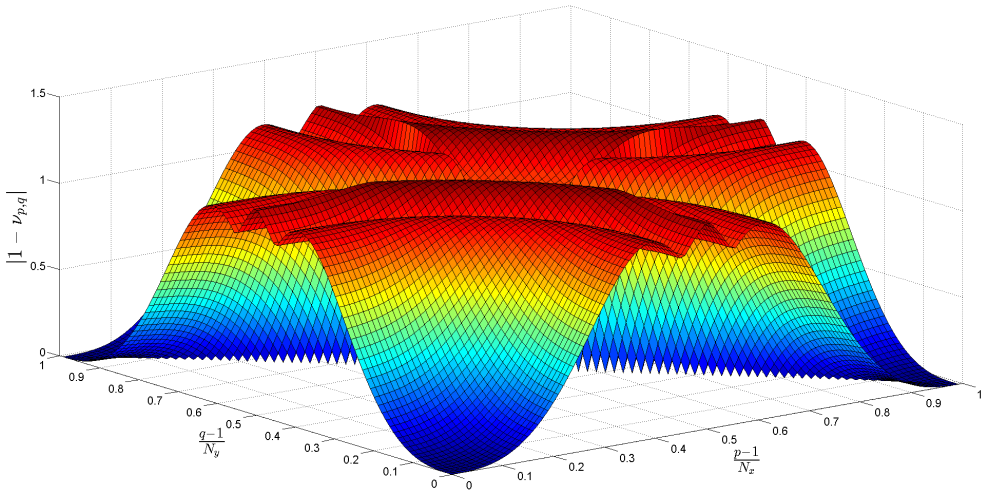
Consider the order of convergence of $|1-\nu_{p,q}|$ to zero with respect to N_x for fixed L ($N_x = N_y$). Using the Upwind and Crank-Nicolson schemes to solve the 2D linear advection problem, we have that $|1-\nu_{1,1}| = 0$. Therefore we need only consider $|1-\nu_{p,q}|$ for $p = 1, \ldots, \frac{N_x+1}{2}$ and $q = 1, \ldots, \frac{N_y+1}{2}$, not both one, in the bound in Equation (5.105).

As we have chosen $N_x = N_y$, the number of points in the plot of $|1-\nu_{p,q}|$ increases in both the x and y-directions by the same factor when N_x is increased. When considering a fixed (p,q) for $p \neq 1$ and $q \neq 1$, the corresponding value of $|1-\nu_{p,q}|$ moves along the x and y-axis when N_x is increased, in the same fashion that $|1-\nu_p|$ did as N_x was increased in Section 4.3.1. When considering a fixed $(p,1)$ for $p \neq 1$ or fixed $(p,1)$ for $q \neq 1$, the corresponding value of $|1-\nu_{p,q}|$ moves along the x-axis or y-axis only respectively.

The Upwind scheme for solving the 2D linear advection problem is a numerically dissipative and dispersive scheme with respect to the resolvable wavenumber components of the numerical solution when $h_1 = h_2 = \frac{1}{2}$ and $N_x = N_y$, as it possesses eigenvalues with this property. However when $p = q$ the scheme also possesses eigenvalues which are numerically non-dissipative and non-dispersive with respect to the resolvable wavenumber components of the numerical solution. This does not change the classification of the scheme due to its numerically dissipative and dispersive prop-



(a) A surface plot for $|1-\nu_{p,q}|$ for the Upwind scheme in (5.29).



(b) A surface plot for $|1-\nu_{p,q}|$ for the Crank-Nicolson scheme in (5.30).

Figure 5.6: *The values of* $|1-\nu_{p,q}|$ *plotted against the corresponding normalised wavenumber in each direction ie:* $\frac{p-1}{N_x}$ *and* $\frac{q-1}{N_y}$*, for* $p=1,\ldots,N_x$ *and* $q=1,\ldots,N_y$*. The schemes considered are the 2D Upwind and 2D Crank-Nicolson schemes for solving the 2D linear advection problem in (5.1), using* $L=4$*,* $N_x=N_y=101$*,* $\mu_1=\mu_2=1$ *and* $h_1=h_2=\frac{1}{2}$ $\left(\Delta t=\frac{1}{2\cdot 101}\right)$*.*

erties, but results in $|1-\nu_{p,q}|=0$ under these conditions. As N_x is increased, $|1-\nu_{p,p}|$ remains zero for some fixed p, $p=1,\ldots,N_x$.

The Crank-Nicolson scheme has a similar property when $h_1=h_2=\frac{1}{2}$ and $N_x=N_y$. The scheme is numerically non-dissipative and dispersive with respect to the resolvable wavenumber components of the numerical solution to the 2D linear advection problem. However when $p+q-2=N_x$, the eigenvalue $\lambda_{p,q}$ of the scheme is numerically non-dissipative and non-dispersive resulting in $|1-\nu_{p,q}|=0$ when $p+q-2=N_x$. Suppose we consider some fixed (p,q) such that $p+q-2=N_x$, then $|1-\nu_{p,q}|=0$. However unlike the Upwind scheme, when N_x is increased, $|1-\nu_{p,q}|$ does not remain zero.

The order of convergence of $|1-\nu_{p,q}|$ to zero for the Upwind and Crank-Nicolson schemes with respect to N_x (as $N_x=N_y$), was identified numerically for fixed (p,q)

using $N_x = 3^\gamma$ and $L = 4$, for $\gamma = 1, \ldots, 7$. The orders of convergence were found to be variable except when p and q, not both one, were both small in comparison to N_x ($\frac{p-1}{N_x} \ll 1$ and $\frac{q-1}{N_y} \ll 1$) where they remained constant; $|1 - \nu_{p,q}| = \mathcal{O}(N_x^{-2})$ for the Upwind scheme (when $p \neq q$) and $|1 - \nu_{p,q}| = \mathcal{O}(N_x^{-3})$ for the Crank-Nicolson scheme. These orders of convergence are determined by the gradient of the surfaces in Figure 5.6 for small p and q, which is dictated by L.

The order of convergence with respect to L

The order of convergence of $|1-\nu_{p,q}|$ to zero with respect to L is investigated numerically similarly to our investigation with respect to N_x. Here we consider fixed (p, q) for $N_x = N_y = 3^7$ and $L = 2^\beta$ where $\beta = 0, \ldots, 7$. The orders of convergence were found to be variable except when p and q not both one, were both small in comparison to N_x ($\frac{p-1}{N_x} \ll 1$ and $\frac{q-1}{N_y} \ll 1$) where they remained constant; $|1 - \nu_{p,q}| = \mathcal{O}(L^2)$ for the Upwind scheme ($p \neq q$) and the Crank-Nicolson scheme.

5.11.2 Asymptotic expansions of $|1 - \nu_{p,q}|$

In this Section, we aim to create asymptotic expansions for $|1 - \nu_{p,q}|$ so that the behaviour of the summations R_1, R_2 and R_3 can be described analytically. When $h_1 = h_2 = \frac{1}{2}$, the eigenvalues of the Upwind scheme are numerically dissipative and dispersive except when $p = q$ when the eigenvalues are numerically non-dissipative and non-dispersive. Therefore we only require an asymptotic expansion of $|1 - \nu_{p,q}|$ for this scheme when $p \neq q$. When $h_1 = h_2 = \frac{1}{2}$, the eigenvalues of the Crank-Nicolson scheme are numerically non-dissipative and dispersive, except when $p = q = 1$ when the eigenvalue is numerically non-dissipative and non-dispersive. Therefore we require an asymptotic expansion of $|1 - \nu_{p,q}|$ when p and q are not both one.

As discussed in Section 4.3.2, asymptotically expanding the form that $|1 - \nu_{p,q}|$ takes when $\lambda_{p,q}$ is a numerically dispersive eigenvalue with respect to its corresponding resolvable wavenumber component, irrespective of its numerically dissipative properties with respect to the resolvable wavenumber components, is challenging. Therefore we are unable to conduct an asymptotic expansion of $|1 - \nu_{p,q}|$ using a Taylor expansion, as was done in Section 4.3.2, for these types of eigenvalue in both considered schemes. Instead we use the knowledge gained from the asymptotic expansions of $|1 - \nu_p|$ in Section 4.3.2 to formulate a possible form for the asymptotic expansion of $|1 - \nu_{p,q}|$ with respect to N_x, N_y and L.

Similarly to the asymptotic expansion in Section 4.3.2, we wish to asymptotically expand a continuous version of $|1-\nu_{p,q}|$. Let $L > 0$ and define $z_1 = \frac{p-1}{N_x}$ and $z_2 = \frac{q-1}{N_y}$. If we consider fixed (p, q) for $p \neq q$ as $N_x, N_y \to \infty$, then $z_1, z_2 \to 0$, so we consider z_1 and z_2 as continuous variables and consider a possible Taylor expansion for $|1 - \nu(z_1, z_2)|$ about $z_1 = z_2 = 0$ to understand the behaviour of $|1 - \nu_{p,q}|$ as $N_x, N_y \to \infty$. The numerical orders of convergence for $|1-\nu_{p,q}|$ to zero with respect to N_x for the Upwind

scheme, when p and q are not equal and are small in comparison to N_x, reveals some interesting clues for the form of the Taylor expansion of $|1-\nu(z_1,z_2)|$ when $z_1 \neq z_2$. We found that $|1-\nu_{p,q}| = \mathcal{O}(L^2 N_x^{-2})$ numerically for p and q small in comparison to N_x, $p \neq q$. This indicates that the Taylor expansion of $|1-\nu(z_1,z_2)|$ is most likely second order. Therefore we consider a Taylor expansion about $z_1 = z_2 = 0$ of the form,

$$|1-\nu(z_1,z_2)| = K_{1a}L^2 z_1^2 + K_{2a}L^2 z_1 z_2 + K_{3a}L^2 z_2^2 + \mathcal{O}(z_1^3 + z_1^2 z_2 + z_1 z_2^2 + z_2^3), \quad (5.114)$$

where $K_{1a}, K_{2a}, K_{3a} \in \mathbb{R}$. This results in,

$$\begin{aligned} |1-\nu_{p,q}| &= K_{1a}L^2\left(\frac{p-1}{N_x}\right)^2 + K_{2a}L^2\left(\frac{p-1}{N_x}\right)\left(\frac{q-1}{N_y}\right) + K_{3a}L^2\left(\frac{q-1}{N_y}\right)^2 \\ &\quad + \mathcal{O}\left(\left[\frac{p-1}{N_x}\right]^3\right) + \mathcal{O}\left(\left[\frac{p-1}{N_x}\right]^2\left[\frac{q-1}{N_y}\right]\right) \\ &\quad + \mathcal{O}\left(\left[\frac{p-1}{N_x}\right]\left[\frac{q-1}{N_y}\right]^2\right) + \mathcal{O}\left(\left[\frac{q-1}{N_y}\right]^3\right), \end{aligned} \quad (5.115)$$

for the 2D Upwind scheme. If we consider $p = 1$ and $q \neq 1$, Equation (5.115) becomes,

$$|1-\nu_{1,q}| = K_{3a}L^2\left(\frac{q-1}{N_y}\right)^2 + \mathcal{O}\left(\left[\frac{q-1}{N_y}\right]^3\right), \quad (5.116)$$

and represents the behaviour of $|1-\nu_{1,q}|$ found numerically when $q \neq 1$ is small in comparison to $N_x = N_y$. Similar is true if we consider Equation (5.115) when $p \neq 1$ and $q = 1$. However it does not represent that $|1-\nu_{p,q}| = 0$ when $p = q$, so when $p = q$ we will not consider this asymptotic expansion and use $|1-\nu_{p,p}| = 0$. Despite this, Equation (5.116) appears to be a good representation of $|1-\nu_{p,q}|$ when $p \neq q$. Therefore we consider,

$$\begin{aligned} &|1-\nu_{p,q}| \\ &\sim \begin{cases} K_{1a}L^2\left(\frac{p-1}{N_x}\right)^2 + K_{2a}L^2\left(\frac{p-1}{N_x}\right)\left(\frac{q-1}{N_y}\right) + K_{3a}L^2\left(\frac{q-1}{N_y}\right)^2, & \text{for } p \neq q, \\ 0, & \text{for } p = q, \end{cases} \\ &\text{as } N_x, N_y \to \infty, \end{aligned} \quad (5.117)$$

for the Upwind scheme and trial the use of,

$$\begin{aligned} &|1-\nu_{p,q}| \\ &\approx \begin{cases} K_{1a}L^2\left(\frac{p-1}{N_x}\right)^2 + K_{2a}L^2\left(\frac{p-1}{N_x}\right)\left(\frac{q-1}{N_y}\right) + K_{3a}L^2\left(\frac{q-1}{N_y}\right)^2, & \text{for } p \neq q, \\ 0, & \text{for } p = q, \end{cases} \end{aligned} \quad (5.118)$$

for the Upwind scheme. Using similar arguments we trial the use of,

$$|1-\nu_{p,q}| \approx K_{1b}L^2\left(\frac{p-1}{N_x}\right)^3 + K_{2b}L^2\left(\frac{p-1}{N_x}\right)^2\left(\frac{q-1}{N_y}\right) + K_{3b}L^2\left(\frac{p-1}{N_x}\right)\left(\frac{q-1}{N_y}\right)^2 + K_{4b}L^2\left(\frac{q-1}{N_y}\right)^3. \quad (5.119)$$

for the Crank-Nicolson scheme, where $K_{1b}, K_{2b}, K_{3b}, K_{4b} \in \mathbb{R}$. In this instance $|1-\nu_{p,q}|$ does not equal zero except when $p = q = 1$, which this asymptotic expansion accommodates.

When considering $h_1 = h_2$ and $N_x = N_y$ as we are here, the Upwind and Crank-Nicolson schemes both have the property that $\lambda_{p,q} = \lambda_{q,p}$ for all $p, q = 1, \ldots, N_x$. Therefore the asymptotic expansion for the Upwind scheme has the property that $K_{1a} = K_{3a}$ and the asymptotic expansion for the Crank-Nicolson scheme has the property that $K_{1b} = K_{4b}$ and $K_{2b} = K_{3b}$.

5.11.3 Analysis of the summations comprising the bound on the error in the analysis vector

In this Section, the orders of convergence for the summations R_1, R_2 and R_3 are identified numerically with respect to N_x, under the conditions set out in Section 5.11.1 for identifying the order of convergence for $|1-\nu_{p,q}|$ numerically with respect to N_x. We only investigate the convergence of R_1, R_2 and R_3 to zero with respect to N_x here. The results of Section 4.3.3 showed that when investigating the order of convergence of S_1 to S_6 to zero, their behaviour with respect to N_x was well understood, whilst their behaviour with respect to L was not. The form of R_1, R_2 and R_3 is similar to that of S_1 to S_6, therefore it is reasonable to assume that the behaviour of R_1, R_2 and R_3 might be well understood with respect to N_x but not L. Until the behaviour of S_1 to S_6 with respect to L can be understood, we leave the behaviour of R_1, R_2 and R_3 with respect to L as future work.

The order of convergence of R_1

$$R_1 = N_x N_y \sum_{p=2}^{\frac{N_x+1}{2}} \sum_{q=2}^{\frac{N_y+1}{2}} \frac{|1-\nu_{p,q}|^2}{|p-1|^{2(r_1+1)}|q-1|^{2(r_2+1)}}$$

This summation is composed from the amplification factors and the bound on the 2D Fourier coefficients of $u_0(x,y)$, giving it an explicit dependence on the regularities of $u_0(x,y)$ in the x-direction (r_1) and the y-direction (r_2). The orders of convergence for R_1 to zero with respect to N_x (as $N_x = N_y$) found numerically for the Upwind scheme, are given in Table 5.3.

The results in Table 5.3 for the Upwind scheme for large r_1 ($r_1 \gg 1$) and $r_2 =$

$0, \ldots, 7$ were identified by considering,

$$R_1 \sim N_x N_y \sum_{q=3}^{\frac{N_y+1}{2}} \frac{|1-\nu_{2,q}|^2}{|q-1|^{2(r_2+1)}}, \tag{5.120}$$

as $|1-\nu_{2,2}| = 0$. Similarly when $r_1 = 0, \ldots, 7$ and r_2 is large ($r_2 \gg 1$), the orders of convergence in Table 5.3 are calculated by considering,

$$R_1 \sim N_x N_y \sum_{p=3}^{\frac{N_x+1}{2}} \frac{|1-\nu_{p,2}|^2}{|p-1|^{2(r_1+1)}}, \tag{5.121}$$

as $|1-\nu_{2,2}| = 0$. When r_1 and r_2 are both large ($r_1 \gg 1$ and $r_2 \gg 1$) we consider,

$$R_1 \sim N_x N_y \left\{ \frac{|1-\nu_{2,3}|^2}{2^{2(r_1+1)}} + \frac{|1-\nu_{3,2}|^2}{2^{2(r_2+1)}} \right\}, \tag{5.122}$$

as $|1-\nu_{2,2}| = 0$.

Examining the numerical orders of convergence for R_1 to zero with respect to N_x in Table 5.3 for the Upwind scheme when $h_1 = h_2 = \frac{1}{2}$, we see that it is the minimum of the regularities in the x- and y-directions that determine the order of convergence. This property is seen in the analytical orders of convergence in Equation (5.126). This property would be expected when either r_1 or r_2 is zero as if $r_1 = 0$ and $r_2 \in \mathbb{N}$, Gibb's phenomenon is present in the x-direction but not in the y-direction so we would not expect the order of convergence with respect to N_x to be improved when compared to the same when $r_1 = r_2 = 0$. We also see that R_1 does not decay to zero when either r_1 or r_2 is zero, showing that the error increases. Once the regularities in the x- and y-directions are such that $r_1 \geq 2$ and $r_2 \geq 2$, the order of convergence of R_1 to zero with respect to N_x saturates at $\mathcal{O}(N_x^{-2})$. Similar properties are seen in the numerical orders of convergence for R_1 to zero with respect to N_x for the Crank-Nicolson scheme when $h_1 = h_2 = \frac{1}{2}$, given in Table 5.4.

The results in Table 5.4 for the Crank-Nicolson scheme for large r_1 ($r_1 \gg 1$) and $r_2 = 0, \ldots, 7$ were identified by considering,

$$R_1 \sim N_x N_y \sum_{q=2}^{\frac{N_y+1}{2}} \frac{|1-\nu_{2,q}|^2}{|q-1|^{2(r_2+1)}}. \tag{5.123}$$

Similarly, when $r_1 = 0, \ldots, 7$ and r_2 is large ($r_2 \gg 1$), the order of convergence in Table 5.4 are calculated by considering,

$$R_1 \sim N_x N_y \sum_{p=2}^{\frac{N_x+1}{2}} \frac{|1-\nu_{p,2}|^2}{|p-1|^{2(r_1+1)}}. \tag{5.124}$$

When r_1 and r_2 are both large ($r_1 \gg 1$ and $r_2 \gg 1$) we consider, When $r_1, r_2 \to \infty$,

$$R_1 \sim N_x N_y |1 - \nu_{2,2}|^2. \tag{5.125}$$

The analytical orders of convergence for R_1 to zero with respect to N_x for the Crank-Nicolson scheme when $h_1 = h_2 = \frac{1}{2}$, are given in Equation (5.127). These match the numerical orders of convergence given for varying regularities in Table 5.4. The order of convergence for the Crank-Nicolson scheme is seen to saturate at $\mathcal{O}(N_x^{-4})$ when $r_1 \geq 3$ and $r_2 \geq 3$. Comparing this with the saturation point of the Upwind scheme, we see that the regularities at the saturation point are larger for the Crank-Nicolson scheme and the error in R_1 decays at a faster rate.

α for R_1 when considering the Upwind scheme.									
r_1 \ r_2	0	1	2	3	4	5	6	7	$r_2 \gg 1$
0	1.0232	1.0090	1.0083	1.0082	1.0082	1.0082	1.0082	1.0082	1.0082
1	1.0090	-9.4741×10^{-1}	-9.5082×10^{-1}	-9.5063×10^{-1}	-9.5071×10^{-1}	-9.5073×10^{-1}	-9.5073×10^{-1}	-9.5073×10^{-1}	-9.5073×10^{-1}
2	1.0083	-9.5087×10^{-1}	-1.9919	-1.9914	-1.9910	-1.9911	-1.9911	-1.9912	-1.9912
3	1.0082	-9.5069×10^{-1}	-2.0036	-2.0276	-2.0201	-2.0202	-2.0206	-2.0207	-2.0208
4	1.0082	-9.5077×10^{-1}	-2.0046	-2.0369	-2.0190	-2.0141	-2.0145	-2.0148	-2.0150
5	1.0082	-9.5079×10^{-1}	-2.0049	-2.0402	-2.0213	-2.0093	-2.0073	-2.0076	-2.0080
6	1.0082	-9.5080×10^{-1}	-2.0050	-2.0411	-2.0233	-2.0095	-2.0042	-2.0034	-2.0038
7	1.0082	-9.5080×10^{-1}	-2.0050	-2.0414	-2.0240	-2.0104	-2.0040	-2.0018	-2.0017
$r_1 \gg 1$	1.0082	-9.5080×10^{-1}	-2.0050	-2.0414	-2.0242	-2.0109	-2.0046	-2.0020	-2.0000

Table 5.3: *The numerical orders of convergence to zero with respect to N_x for $R_1 = \mathcal{O}(N_x^{\alpha})$ using the Upwind scheme, given to 4dp (decimal places) for $h_1 = h_2 = \frac{1}{2}$ and $\mu_1 = \mu_2 = 1$. The results were identified using fixed $L = 4$ and by considering N_x in the form $N_x = 3^{\gamma}$ $\left(\Delta t = \frac{1}{2N_x}\right)$, where $\gamma = 1, \ldots, 7$. The results displayed here are the orders of convergence for the largest values of N_x considered. The full set of results can be found in Appendix C.1.1, in Tables C.1-C.9.*

If R_1 is considered analytically when $N_x = N_y$, $h_1 = h_2 = \frac{1}{2}$ and $\mu_1 = \mu_2 = 1$, then the calculations in Appendix C.1.2 show that by substituting Equation 5.118 into R_1 we obtain,

$$R_1 = \begin{cases} \mathcal{O}\left(L^2 N_x\right), & \text{for } \min(r_1, r_2) = 0, \\ \mathcal{O}\left(L^2 N_x^{-1}\right), & \text{for } \min(r_1, r_2) = 1, \\ \mathcal{O}\left(L^2 N_x^{-2}\right), & \text{for } \min(r_1, r_2) \geq 2. \end{cases} \tag{5.126}$$

α for R_1 when considering the Crank-Nicolson scheme.									
r_1 \ r_2	0	1	2	3	4	5	6	7	$r_2 \gg 1$
0	9.9914×10^{-1}	9.9951×10^{-1}	9.9957×10^{-1}	9.9958×10^{-1}	9.9958×10^{-1}	9.9958×10^{-1}	9.9958×10^{-1}	9.9958×10^{-1}	9.9958×10^{-1}
1	9.9951×10^{-1}	-1.0013	-1.0012	-1.0011	-1.0011	-1.0011	-1.0011	-1.0011	-1.0011
2	9.9957×10^{-1}	-1.0012	-3.0168	-3.0295	-3.0272	-3.0269	-3.0268	-3.0268	-3.0268
3	9.9958×10^{-1}	-1.0011	-3.0295	-3.9990	-3.9994	-3.9994	-3.9994	-3.9994	-3.9994
4	9.9958×10^{-1}	-1.0011	-3.0272	-3.9994	-4.0000	-4.0000	-4.0000	-4.0000	-4.0000
5	9.9958×10^{-1}	-1.0011	-3.0269	-3.9994	-4.0000	-4.0000	-4.0000	-4.0000	-4.0000
6	9.9958×10^{-1}	-1.0011	-3.0268	-3.9994	-4.0000	-4.0000	-4.0000	-4.0000	-4.0000
7	9.9958×10^{-1}	-1.0011	-3.0268	-3.9994	-4.0000	-4.0000	-4.0000	-4.0000	-4.0000
$r_1 \gg 1$	9.9958×10^{-1}	-1.0011	-3.0268	-3.9994	-4.0000	-4.0000	-4.0000	-4.0000	-4.0000

Table 5.4: *The numerical orders of convergence to zero with respect to N_x for $R_1 = \mathcal{O}(N_x^{\alpha})$ using the Crank-Nicolson scheme, given to 4dp (decimal places) for $h_1 = h_2 = \frac{1}{2}$ and $\mu_1 = \mu_2 = 1$. The results were identified using fixed $L = 4$ and by considering N_x in the form $N_x = 3^{\gamma}$ $\left(\Delta t = \frac{1}{2N_x}\right)$, where $\gamma = 1, \ldots, 7$. The results displayed here are the orders of convergence for the largest values of N_x considered. The full set of results can be found in Appendix C.1.3, in Tables C.10-C.18.*

If R_1 is considered analytically when $N_x = N_y$, $h_1 = h_2 = \frac{1}{2}$ and $\mu_1 = \mu_2 = 1$, then the calculations in Appendix C.1.4 show that by substituting Equation 5.119 into R_1 we obtain,

$$R_1 = \begin{cases} \mathcal{O}(L^2 N_x), & \text{for } \min(r_1, r_2) = 0, \\ \mathcal{O}(L^2 N_x^{-1}), & \text{for } \min(r_1, r_2) = 1, \\ \mathcal{O}(L^2 N_x^{-3}), & \text{for } \min(r_1, r_2) = 2, \\ \mathcal{O}(L^2 N_x^{-4}), & \text{for } \min(r_1, r_2) \geq 3. \end{cases} \tag{5.127}$$

The order of convergence of R_2

$$R_2 = N_x N_y \sum_{p=2}^{\frac{N_x+1}{2}} \frac{|1-\nu_{p,1}|^2}{|p-1|^{2(r_1+1)}}$$

This summation is also composed from the amplification factors and bound on the 2D Fourier coefficients of $u_0(x,y)$. However as $q=1$, it only has an explicit dependence on the regularity of $u_0(x,y)$ in the x-direction (r_1). The orders of convergence for R_2 to zero with respect to N_x (as $N_x = N_y$) found numerically for the Upwind scheme, are given in Table 5.5.

The results in Table 5.5 for the Upwind scheme and in Table 5.6 for the Crank-Nicolson scheme for large r_1 ($r_1 \gg 1$) were identified by considering,

$$R_2 \sim N_x N_y |1-\nu_{2,1}|^2. \tag{5.128}$$

Examining the orders of convergence for R_2 to zero with respect to N_x in Table 5.5 for the Upwind scheme when $h_1 = h_2 = \frac{1}{2}$, we see that the error in R_2 increases as N_x is increased when $r_1 = 0$. Despite R_2 only being dependent on r_1 and having a similar form to S_1, R_2 increases as N_x is increased when $r_1 = 0$, similarly to R_1 when either $r_1 = 0$ or $r_2 = 0$. When considering higher regularities, R_2 decays to zero as N_x is increased. The order of convergence of R_2 to zero saturates at $\mathcal{O}(N_x^{-2})$ when $r_1 \geq 2$ for the Upwind scheme. Similar properties are seen in the numerical orders of convergence for R_2 to zero with respect to N_x for the Crank-Nicolson scheme when $h_1 = h_2 = \frac{1}{2}$. These results are given in Table 5.6. The analytical orders of convergence for R_2 to zero with respect to N_x for the Upwind scheme when $h_1 = h_2 = \frac{1}{2}$, are given by Equation (5.129). These match the numerical orders of convergence given for varying regularities in Table 5.5.

The analytical orders of convergence for R_2 to zero with respect to N_x for the Crank-Nicolson scheme when $h_1 = h_2 = \frac{1}{2}$, are given by Equation (5.130). These match the numerical orders of convergence given for varying regularities in Table 5.6. The order of convergence of R_2 to zero for the Crank-Nicolson scheme saturates at $\mathcal{O}(N_x^{-4})$ when $r_1 \geq 3$. If we again compare this to the similar saturation point of R_2 for the Upwind scheme, we see that saturation point is achieved at a higher regularity for the Crank-Nicolson scheme. The rate of decay of R_2 to zero at saturation point as N_x is increased, is much faster for the Crank-Nicolson scheme.

α for R_2 when considering the Upwind scheme.									
r_1 \ r_2	0	1	2	3	4	5	6	7	$r_2 \gg 1$
0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1	-9.9475×10^{-1}	-9.9475×10^{-1}	-9.9475×10^{-1}	-9.9475×10^{-1}	-9.9475×10^{-1}	-9.9475×10^{-1}	-9.9475×10^{-1}	-9.9475×10^{-1}	-9.9475×10^{-1}
2	−1.9529	−1.9529	−1.9529	−1.9529	−1.9529	−1.9529	−1.9529	−1.9529	−1.9529
3	−1.9396	−1.9396	−1.9396	−1.9396	−1.9396	−1.9396	−1.9396	−1.9396	−1.9396
4	−1.9358	−1.9358	−1.9358	−1.9358	−1.9358	−1.9358	−1.9358	−1.9358	−1.9358
5	−1.9350	−1.9350	−1.9350	−1.9350	−1.9350	−1.9350	−1.9350	−1.9350	−1.9350
6	−1.9348	−1.9348	−1.9348	−1.9348	−1.9348	−1.9348	−1.9348	−1.9348	−1.9348
7	−1.9348	−1.9348	−1.9348	−1.9348	−1.9348	−1.9348	−1.9348	−1.9348	−1.9348
$r_1 \gg 1$	−1.9347	−1.9347	−1.9347	−1.9347	−1.9347	−1.9347	−1.9347	−1.9347	−1.9347

Table 5.5: *The numerical orders of convergence to zero with respect to N_x for $R_2 = \mathcal{O}(N_x^{\alpha})$ using the Upwind scheme, given to 4dp (decimal places) for $h_1 = h_2 = \frac{1}{2}$ and $\mu_1 = \mu_2 = 1$. The results were identified using fixed $L = 4$ and by considering N_x in the form $N_x = 3^{\gamma}$ $\left(\Delta t = \frac{1}{2N_x}\right)$, where $\gamma = 1, \ldots, 7$. The results displayed here are the orders of convergence created using $N_x = 3^5$ and $N_x = 3^6$. See Remark 5.12. The full set of results can be found in Appendix C.2.1, in Table C.19.*

If R_2 is considered analytically when $N_x = N_y$, $h_1 = h_2 = \frac{1}{2}$ and $\mu_1 = \mu_2 = 1$, then the calculations in Appendix C.2.2 show that by substituting Equation 5.118 into R_2 we obtain,

$$R_2 = \begin{cases} \mathcal{O}(L^2 N_x), & \text{for } r_1 = 0, \\ \mathcal{O}(L^2 N_x^{-1}), & \text{for } r_1 = 1, \\ \mathcal{O}(L^2 N_x^{-2}), & \text{for } r_1 \geq 2, \end{cases} \tag{5.129}$$

for all $r_2 \in \mathbb{N}_0$.

α for R_2 when considering the Crank-Nicolson scheme.									
r_1 \ r_2	0	1	2	3	4	5	6	7	$r_2 \gg 1$
0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1	−1.0000	−1.0000	−1.0000	−1.0000	−1.0000	−1.0000	−1.0000	−1.0000	−1.0000
2	−2.9985	−2.9985	−2.9985	−2.9985	−2.9985	−2.9985	−2.9985	−2.9985	−2.9985
3	−3.9974	−3.9974	−3.9974	−3.9974	−3.9974	−3.9974	−3.9974	−3.9974	−3.9974
4	−4.0000	−4.0000	−4.0000	−4.0000	−4.0000	−4.0000	−4.0000	−4.0000	−4.0000
5	−4.0000	−4.0000	−4.0000	−4.0000	−4.0000	−4.0000	−4.0000	−4.0000	−4.0000
6	−4.0000	−4.0000	−4.0000	−4.0000	−4.0000	−4.0000	−4.0000	−4.0000	−4.0000
7	−4.0000	−4.0000	−4.0000	−4.0000	−4.0000	−4.0000	−4.0000	−4.0000	−4.0000
$r_1 \gg 1$	−4.0000	−4.0000	−4.0000	−4.0000	−4.0000	−4.0000	−4.0000	−4.0000	−4.0000

Table 5.6: *The numerical orders of convergence to zero with respect to N_x for $R_2 = \mathcal{O}(N_x^{\alpha})$ using the Crank-Nicolson scheme, given to 4dp (decimal places) for $h_1 = h_2 = \frac{1}{2}$ and $\mu_1 = \mu_2 = 1$. The results were identified using fixed $L = 4$ and by considering N_x in the form $N_x = 3^{\gamma}$ $\left(\Delta t = \frac{1}{2N_x}\right)$, where $\gamma = 1, \ldots, 7$. The results displayed here are the orders of convergence for the largest values of N_x considered. The full set of results can be found in Appendix C.2.3, in Table C.20.*

If R_2 is considered analytically when $N_x = N_y$, $h_1 = h_2 = \frac{1}{2}$ and $\mu_1 = \mu_2 = 1$, then the calculations in Appendix C.1.4 show that by substituting Equation 5.119 into R_2 we obtain,

$$R_2 = \begin{cases} \mathcal{O}(L^2 N_x), & \text{for } r_1 = 0, \\ \mathcal{O}(L^2 N_x^{-1}), & \text{for } r_1 = 1, \\ \mathcal{O}(L^2 N_x^{-3}), & \text{for } r_1 = 2, \\ \mathcal{O}(L^2 N_x^{-4}), & \text{for } r_1 \geq 3, \end{cases} \tag{5.130}$$

for all $r_2 \in \mathbb{N}_0$.

The order of convergence of R_3

$$R_3 = N_x N_y \sum_{q=2}^{\frac{N_y+1}{2}} \frac{|1-\nu_{1,q}|^2}{|q-1|^{2(r_2+1)}}$$

This summation is composed from the amplification factors and bound on the 2D Fourier coefficients of $u_0(x,y)$. However as $p=1$, it only has an explicit dependence on the regularity of $u_0(x,y)$ in the y-direction (r_2). As $N_x = N_y$ and $h_1 = h_2$ when identifying the numerical orders of convergence of R_3 to zero, this results in $\lambda_{p,q} = \lambda_{q,p}$ for both the Upwind and the Crank-Nicolson scheme. In this instance, this allows R_3 to re-written as,

$$R_3 = N_x^2 \sum_{q=2}^{\frac{N_x+1}{2}} \frac{|1-\nu_{q,1}|^2}{|q-1|^{2(r_2+1)}},$$

which is almost identical to R_2 under the same conditions, but with the dependence on r_1 replaced by a dependence on r_2. Consequently, the numerical orders of convergence for R_3 for the Upwind and Crank-Nicolson schemes in Tables 5.7 and 5.8 respectively, are identical to those for R_2 in Tables 5.5 and 5.6 respectively but with the role of r_1 swapped for r_2. This can be seen by comparing the corresponding Tables. Therefore the comments on the numerical results for R_2 are applicable to R_3 but with r_2 replacing r_1.

By similar reasoning, when r_2 is large ($r_2 \gg 1$), the numerical order of convergence for R_3 in Tables 5.7 and 5.8 are identified by,

$$R_3 \sim N_x N_y |1-\nu_{1,2}|^2. \tag{5.131}$$

We can also see that the analytic orders of convergence for R_3 for the Upwind and Crank-Nicolson schemes in Equations (5.132) and (5.133) respectively, match their respective numerical orders of convergence for R_3 with respect to N_x.

α for R_3 when considering the Upwind scheme.									
r_1 \ r_2	0	1	2	3	4	5	6	7	$r_2 \gg 1$
0	1.0000	-9.9475×10^{-1}	-1.9529	-1.9396	-1.9358	-1.9350	-1.9348	-1.9348	-1.9347
1	1.0000	-9.9475×10^{-1}	-1.9529	-1.9396	-1.9358	-1.9350	-1.9348	-1.9348	-1.9347
2	1.0000	-9.9475×10^{-1}	-1.9529	-1.9396	-1.9358	-1.9350	-1.9348	-1.9348	-1.9347
3	1.0000	-9.9475×10^{-1}	-1.9529	-1.9396	-1.9358	-1.9350	-1.9348	-1.9348	-1.9347
4	1.0000	-9.9475×10^{-1}	-1.9529	-1.9396	-1.9358	-1.9350	-1.9348	-1.9348	-1.9347
5	1.0000	-9.9475×10^{-1}	-1.9529	-1.9396	-1.9358	-1.9350	-1.9348	-1.9348	-1.9347
6	1.0000	-9.9475×10^{-1}	-1.9529	-1.9396	-1.9358	-1.9350	-1.9348	-1.9348	-1.9347
7	1.0000	-9.9475×10^{-1}	-1.9529	-1.9396	-1.9358	-1.9350	-1.9348	-1.9348	-1.9347
$r_1 \gg 1$	1.0000	-9.9475×10^{-1}	-1.9529	-1.9396	-1.9358	-1.9350	-1.9348	-1.9348	-1.9347

Table 5.7: *The numerical orders of convergence to zero with respect to N_x for $R_3 = \mathcal{O}(N_x^{\alpha})$ using the Upwind scheme, given to 4dp (decimal places) for $h_1 = h_2 = \frac{1}{2}$ and $\mu_1 = \mu_2 = 1$. The results were identified using fixed $L = 4$ and by considering N_x in the form $N_x = 3^{\gamma}$ $\left(\Delta t = \frac{1}{2N_x}\right)$, where $\gamma = 1, \ldots, 7$. The results displayed here are the orders of convergence created using $N_x = 3^5$ and $N_x = 3^6$. See Remark 5.12. The full set of results can be found in Appendix C.3.1, in Table C.21.*

If R_3 is considered analytically when $N_x = N_y$, $h_1 = h_2 = \frac{1}{2}$ and $\mu_1 = \mu_2 = 1$, then the calculations in Appendix C.3.2 show that by substituting Equation 5.118 into R_3 we obtain,

$$R_3 = \begin{cases} \mathcal{O}(L^2 N_x), & \text{for } r_2 = 0, \\ \mathcal{O}(L^2 N_x^{-1}), & \text{for } r_2 = 1, \\ \mathcal{O}(L^2 N_x^{-2}), & \text{for } r_2 \geq 2, \end{cases} \tag{5.132}$$

for all $r_1 \in \mathbb{N}_0$.

α for R_3 when considering the Crank-Nicolson scheme.									
r_1 \ r_2	0	1	2	3	4	5	6	7	$r_2 \gg 1$
0	1.0000	−1.0000	−2.9985	−3.9974	−4.0000	−4.0000	−4.0000	−4.0000	−4.0000
1	1.0000	−1.0000	−2.9985	−3.9974	−4.0000	−4.0000	−4.0000	−4.0000	−4.0000
2	1.0000	−1.0000	−2.9985	−3.9974	−4.0000	−4.0000	−4.0000	−4.0000	−4.0000
3	1.0000	−1.0000	−2.9985	−3.9974	−4.0000	−4.0000	−4.0000	−4.0000	−4.0000
4	1.0000	−1.0000	−2.9985	−3.9974	−4.0000	−4.0000	−4.0000	−4.0000	−4.0000
5	1.0000	−1.0000	−2.9985	−3.9974	−4.0000	−4.0000	−4.0000	−4.0000	−4.0000
6	1.0000	−1.0000	−2.9985	−3.9974	−4.0000	−4.0000	−4.0000	−4.0000	−4.0000
7	1.0000	−1.0000	−2.9985	−3.9974	−4.0000	−4.0000	−4.0000	−4.0000	−4.0000
$r_1 \gg 1$	1.0000	−1.0000	−2.9985	−3.9974	−4.0000	−4.0000	−4.0000	−4.0000	−4.0000

Table 5.8: *The numerical orders of convergence to zero with respect to N_x for $R_3 = \mathcal{O}(N_x^{\alpha})$ using the Crank-Nicolson scheme, given to 4dp (decimal places) for $h_1 = h_2 = \frac{1}{2}$ and $\mu_1 = \mu_2 = 1$. The results were identified using fixed $L = 4$ and by considering N_x in the form $N_x = 3^{\gamma}$ $\left(\Delta t = \frac{1}{2N_x}\right)$, where $\gamma = 1, \ldots, 7$. The results displayed here are the orders of convergence for the largest values of N_x considered. The full set of results can be found in Appendix C.3.3, in Table C.22.*

If R_3 is considered analytically when $N_x = N_y$, $h_1 = h_2 = \frac{1}{2}$ and $\mu_1 = \mu_2 = 1$, then the calculations in Appendix C.3.4 show that by substituting Equation 5.119 into R_3 we obtain,

$$R_3 = \begin{cases} \mathcal{O}(L^2 N_x), & \text{for } r_2 = 0, \\ \mathcal{O}(L^2 N_x^{-1}), & \text{for } r_2 = 1, \\ \mathcal{O}(L^2 N_x^{-3}), & \text{for } r_2 = 2, \\ \mathcal{O}(L^2 N_x^{-4}), & \text{for } r_2 \geq 3, \end{cases} \tag{5.133}$$

for all $r_1 \in \mathbb{N}_0$.

Remark 5.12. *The results displayed in Tables 5.5 and 5.7 are created using $N_x = 3^5$ and $N_x = 3^6$, unlike those in Tables 5.3, 5.4, 5.6 and 5.8, which are created using the largest values of N_x considered ie: $N_x = 3^6$ and $N_x = 3^7$. The results in Tables 5.5 and 5.7 are for R_2 and R_3 respectively for the Upwind scheme. Considering the orders of convergence created for these summations for the Upwind scheme using $N_x = 3^6$ and $N_x = 3^7$, Tables C.19 and C.21 respectively show that the behaviour of these summations has changed and does not match the analytical results in Equations 5.129 and 5.132. However the equivalent results for the Crank-Nicolson scheme in Tables C.20 and C.22 do match their analytical results. This indicates that it is not the Matlab code that is a problem. The values of R_2 and R_3 generated for the Upwind scheme when $N_x = 3^6$ and $N_x = 3^7$ are not small enough for numerical errors to be the reason for the change in behaviour. The change in behaviour is most likely due to some property specific to the Upwind scheme and is left as future work to understand.*

5.11.4 The dominant summation

Now the order of convergence of summations R_1, R_2 and R_3 to zero with respect to N_x (as $N_x = N_y$) have been examined for $h_1 = h_2 = \frac{1}{2}$, we can determine which of these summations is dominant. Comparing their orders of convergence for different regularity initial conditions, we see that the behaviour of R_1 to zero with respect to N_x encompasses the behaviour of R_2 and R_3 to zero with respect to N_x. Therefore in order to determine if R_1 and consequently the bound in Equation (5.105) can be used to represent and hence understand the behaviour of the l_2-norm of the error in the analysis vector for the 2D linear advection problem, we compare the order of convergence of R_1 to zero with the numerical order of convergence of the l_2-norm of the error in the analysis vector to zero, found through 4D-Var data assimilation numerical experiments. The results from numerical strong constraint 4D-Var data assimilation experiments for comparison are detailed in the following Section.

5.12 Results from strong constraint 4D-Var numerical experiments

In this Section, we perform strong constraint 4D-Var data assimilation numerical experiments to identify the behaviour of the l_2-norm of the error in the analysis vector. We investigate the order of convergence of this quantity to zero with respect to N_x using the Upwind, Crank-Nicolson and MNIMC finite difference schemes and several different multiplicatively separable functions for $u_0(x, y) = \hat{u}_1(x)\hat{u}_2(y)$ over $[0, 1) \times [0, 1)$. The initial conditions considered were chosen due to the regularity of their constituent functions $\hat{u}_1(x)$ and $\hat{u}_2(y)$ over $(0, 1)$ and $(0, 1)$, respectively. These functions are:

- The square-square function ($r_1 = 0$, $r_2 = 0$),

$$\hat{u}_1(x) = \begin{cases} -\frac{1}{2}, & \text{for } x \in \left[0, \frac{1}{4}\right) \bigcup \left(\frac{1}{2}, 1\right), \\ \frac{1}{2}, & \text{for } x \in \left[\frac{1}{4}, \frac{1}{2}\right], \end{cases} \tag{5.134}$$

  and the function $\hat{u}_2(y)$ is defined identically to $\hat{u}_1(x)$, but in terms of y. Here, $v_1 = v_2 = v_3 = v_4 = \frac{1}{2}$, $s_1 = s_2 = 3$, $Q_j = 9$ for all $j = 1, \ldots, 4$ and $w_1 = w_2 = 2$ as N_x and N_y are both odd (as discussed in Section 4.3), so $A_1 = \frac{1}{4}$, $A_2 = A_3 = \frac{27}{2\pi}$, $A_4 = \frac{81}{\pi^2}$, $A_5 = A_6 = 1$, $A_7 = A_8 = \frac{3}{\pi}$ and $A_9 = 2$.

- The square-triangle function ($r_1 = 0$, $r_2 = 1$), the function $\hat{u}_1(x)$ is defined as in (5.134) and

$$\hat{u}_2(y) = \begin{cases} -\frac{1}{2}, & \text{for } y \in \left[0, \frac{1}{4}\right) \bigcup \left(\frac{1}{2}, 1\right), \\ 8y - \frac{5}{2}, & \text{for } y \in \left[\frac{1}{4}, \frac{3}{8}\right], \\ -8y + \frac{7}{2}, & \text{for } y \in \left(\frac{3}{8}, \frac{1}{2}\right]. \end{cases} \tag{5.135}$$

  Here, $v_1 = v_2 = v_3 = \frac{1}{2}$, $v_4 = 8$, $s_1 = 3$, $s_2 = 4$, $Q_1 = Q_3 = 9$, $Q_2 = Q_4 = 12$ and $w_1 = 2$ as N_x is odd, so $A_1 = \frac{1}{4}$, $A_2 = \frac{192}{\pi^2}$, $A_3 = \frac{27}{2\pi}$, $A_4 = \frac{1152}{\pi^3}$, $A_5 = 1$ and $A_7 = \frac{32}{\pi^2}$.

- The triangle-triangle function ($r_1 = 1$, $r_2 = 1$), the functions $\hat{u}_1(x)$ and $\hat{u}_2(y)$ are defined as in (5.135). Here, $v_1 = v_3 = \frac{1}{2}$, $v_2 = v_4 = 8$, $s_1 = 3$, $s_2 = 4$ and $Q_j = 1$ for all $j = 1, \ldots, 4$, so $A_1 = \frac{1}{4}$, $A_2 = A_3 = \frac{192}{\pi^2}$ and $A_4 = \frac{1536}{\pi^4}$.

- The 2D Gaussian function $\mathcal{N}\left(\left[\frac{1}{2}, \frac{1}{2}\right]^T, \frac{1}{100} I_2\right)$ ($r_1 \gg 1$, $r_2 \gg 1$),

$$u_0(x, y) = \frac{50}{\pi} e^{-50\left\{\left(x - \frac{1}{2}\right)^2 + \left(y - \frac{1}{2}\right)^2\right\}}, \tag{5.136}$$

  which decomposes into,

$$\hat{u}_1(x) = \sqrt{\frac{50}{\pi}} e^{-50\left(x - \frac{1}{2}\right)^2}, \tag{5.137}$$

$$\hat{u}_2(y) = \sqrt{\frac{50}{\pi}} e^{-50\left(y - \frac{1}{2}\right)^2}. \tag{5.138}$$

  Here $v_1 = v_3 = \sqrt{\frac{50}{\pi}}$. These Gaussian functions need to be considered for r_1 and r_2 sufficiently large but not infinite, as discussed for the Gaussian function in Section 4.3.

It is important to understand the relationship between the regularity of the functions constructing a multiplicatively separable $u_0(x, y)$ and the locations of the discontinuities in $u_0(x, y)$. Consider the 1D square function $\hat{u}_1(x)$ in (5.134). This function has regularity $r_1 = 0$ and results in lines of jump discontinuities running parallel to the y-axis in $u_0(x, y)$. This is counter-intuitive as you might think that $r_1 = 0$ is an indicator of the continuity of $u_0(x, y)$ in the x-direction, when in fact it indicates that jump

discontinuities exist in the y-direction. Similarly when $r_2 = 0$, lines of jump discontinuities are present in $u_0(x, y)$ parallel to the x-axis, resulting in jump discontinuities in the x-direction. As a consequence of these properties, only lines of discontinuity can be present in separable $u_0(x, y)$, so there are no lone points of jump discontinuity. When $r_1, r_2 \in \mathbb{N}$, $u_0(x, y)$ is continuous in both the x- and y-directions.

The strong constraint 4D-Var data assimilation numerical experiments to identify the order of convergence of the l_2-norm of the error in the analysis vector to zero with respect to N_x, were conducted using $N_x = N_y$ and $h_1 = h_2 = \frac{1}{2}$, using the same values for N_x and L as detailed in Section 5.11.1. The challenge associated with investigating the order of convergence in this way, is that the vectors and matrices associated with the problem have dimension N_xN_y and $N_xN_y \times N_xN_y$ respectively, so increasing N_x and N_y simultaneously increases the computational cost of the problem rapidly.

The numerical results for the strong constraint 4D-Var data assimilation numerical experiments, using the multiplicatively separable initial conditions, are plotted for N_x and N_y as powers of three in Figure 5.7. Examining Figure 5.7, we see that the l_2-norm of the error in the analysis vector increases for each considered scheme, for both the square-square and square-triangle initial conditions. The order of convergence of the error to zero with respect to N_x appears from the limited results to be approximately $\mathcal{O}(N_x)$, for each initial condition and scheme. This agrees with the numerical and analytical analysis of the dominant summation R_1 in Section 5.11.3. These initial conditions both contain discontinuities, so the increase in the l_2-norm of the error of the analysis vector may be due to Gibb's phenomenon as we are considering finite N_x and N_y. Gibb's phenomenon diminishes as N_x and N_y are increased, but the error between the continuous Fourier series representation of the numerical solution at the location of discontinuities, is always present. However, the increase in the error is a surprising result as we might have expected the order of convergence to remain constant based on the results from the 1D linear advection problem.

The square-square function contains lines of jump discontinuities parallel to both the x- and y-axis, whilst the square-triangle function contains discontinuities parallel to the y-axis. Despite this difference, the numerical orders of convergence for the l_2-norm of the error in the analysis vector to zero as N_x and N_y are increased (as $N_x = N_y$), are similar. This is surprising as you might expect the lack of discontinuities in one direction for the square-triangle function, to alter the order of convergence. This behaviour needs further investigation.

Figure 5.7 shows that the l_2-norm of the error in the analysis vector decays rapidly for each scheme, when considering the Gaussian initial condition. This may indicate that the results for the 2D linear advection problem are consistent with the results of the 1D linear advection problem, in that the l_2-norm of the error in the analysis vector for initial conditions containing no discontinuities, decay to zero. The results for the triangle-triangle initial condition do not provide a clear order of convergence for the l_2-norm of the error in the analysis vector, to confirm this hypothesis. Larger

values of N_x need to be considered to compare the rate of decay of the error for the triangle-triangle and Gaussian initial conditions, with the order of convergence of R_1 to zero from Section 5.11.3. However Figure 5.7 does show that the error for the Crank-Nicolson scheme decays to zero with respect to N_x faster than for the Upwind scheme, for the Gaussian initial condition. This supports the idea that R_1 may be the dominant summation of the bound on the l_2-norm of the error in the analysis vector for the 2D linear advection problem, as Section 5.11.3 found that when r_1 and r_2 are both large, R_1 decays to zero faster with respect to N_x for the Crank-Nicolson scheme.

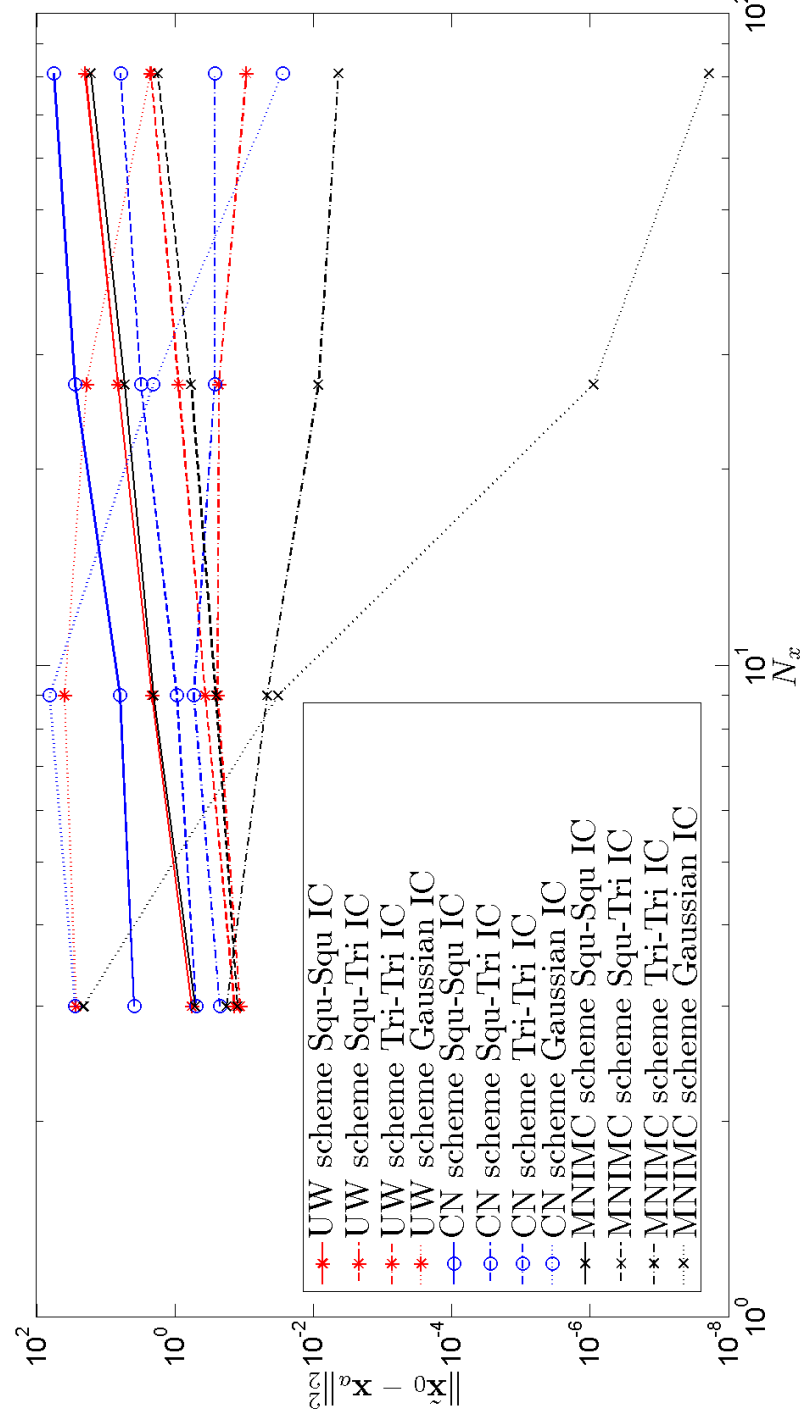


Figure 5.7: *The square of the l_2-norm of the error in the analysis vector, calculated through strong constraint 4D-Var data assimilation numerical experiments, under the influence of errors introduced by finite difference approximations in the forward model. The results were generated using the Upwind (UW), Crank-Nicolson (CN) and MNIMC schemes as the forward models for solving the 2D linear advection problem in (5.1), using $h_1 = h_2 = 0.5$, $\mu_1 = \mu_2 = 1$ and $L = 4$, where $N_x = N_y = 3^{\alpha}$ for $\alpha = 1, \ldots, 4$ ($\Delta t = \frac{1}{2N_x}$). The functions considered for $u_0(x, y)$ in these experiments are defined in Section 5.12 by the multiplicatively separable square-square (squ-squ IC), square-triangle (squ-tri IC), triangle-triangle (tri-tri IC) and the 2D Gaussian (Gaussian IC) functions. The results are plotted using logarithmic scales to demonstrate the order of convergence as both N_x and N_y are increased at the same rate.*

Numerical experiments were also performed using functions that were not multiplicatively separable, but could possibly hold similar regularities once regularity has been defined for non-separable functions. For example, the square function initial condition in (5.139) is discontinuous in the x- and y-directions, so is likely to hold a regularity equivalent to $r_1 = 0$ and $r_2 = 0$ for a multiplicatively separable initial condition. The numerical results from these non-separable functions can be compared with those from the multiplicatively separable functions which possess the regularity that we predict the non-separable function to possess. The aim is to identify if the numerical results from the multiplicatively non-separable functions, are consistent with the results from the separable functions, to help with the development of a definition for the regularity of multiplicatively non-separable functions. The multiplicatively non-separable initial conditions considered are as follows;

- The 2D square function (predicted to be equivalent to $r_1 = 0$, $r_2 = 0$)

$$u_0(x,y) = \begin{cases} \frac{1}{2}, & \text{for } (x,y) \in \left[\frac{1}{4},\frac{1}{2}\right] \times \left[\frac{1}{4},\frac{1}{2}\right], \\ -\frac{1}{2}, & \text{for } (x,y) \in [0,1) \times [0,1) \setminus \left[\frac{1}{4},\frac{1}{2}\right] \times \left[\frac{1}{4},\frac{1}{2}\right]. \end{cases} \tag{5.139}$$

- The tent function (predicted to be equivalent to $r_1 = 1$, $r_2 = 0$),

$$u_0(x,y) = \begin{cases} 8x - \frac{5}{2}, & \text{for } (x,y) \in \left[\frac{1}{4},\frac{3}{8}\right] \times \left[\frac{1}{4},\frac{1}{2}\right], \\ -8x + \frac{7}{2}, & \text{for } (x,y) \in \left(\frac{3}{8},\frac{1}{2}\right] \times \left[\frac{1}{4},\frac{1}{2}\right], \\ -\frac{1}{2}, & \text{for } (x,y) \in [0,1) \times [0,1) \setminus \left[\frac{1}{4},\frac{1}{2}\right] \times \left[\frac{1}{4},\frac{1}{2}\right]. \end{cases} \tag{5.140}$$

- The square-based pyramid function (predicted to be equivalent to $r_1 = 1$, $r_2 = 1$),

$$u_0(x,y) = \begin{cases} 8y - \frac{5}{2}, & \text{for } x \in \left[y, -y + \frac{3}{4}\right] \text{ and } y \in \left[\frac{1}{4},\frac{3}{8}\right], \\ 8x - \frac{5}{2}, & \text{for } x \in \left[\frac{1}{4},\frac{3}{8}\right) \text{ and } y \in \left(x, -x + \frac{3}{4}\right], \\ -8y + \frac{7}{2}, & \text{for } x \in \left(-y + \frac{3}{4}, y\right) \text{ and } y \in \left(\frac{3}{8},\frac{1}{2}\right], \\ -8x + \frac{7}{2}, & \text{for } x \in \left(\frac{3}{8},\frac{1}{2}\right] \text{ and } y \in \left(-x + \frac{3}{4}, x\right], \\ -\frac{1}{2}, & \text{for } (x,y) \in [0,1) \times [0,1) \setminus \left[\frac{1}{4},\frac{1}{2}\right] \times \left[\frac{1}{4},\frac{1}{2}\right]. \end{cases} \tag{5.141}$$

The numerical results for the strong constraint 4D-Var data assimilation experiments, using these multiplicatively non-separable functions, are plotted for N_x and N_y as powers of three in Figure 5.8.

Examining Figure 5.8, we see that the 2D square and tent initial conditions, which both contain discontinuities, appear to display an increase in the l_2-norm of the error in the analysis vector as N_x is increased. This is consistent with our results from Figure 5.7 for discontinuous initial conditions and the behaviour of R_1 with respect to N_x when $\min(r_1, r_2) = 0$. This may be due to these initial conditions both containing lines of jump discontinuities parallel to the x- and y-axis also. It would be interesting to consider the results from an initial condition containing a lone point of jump discontinuity or one where the line of discontinuity is not parallel to an axis.

The l_2-norm of the error in the analysis vector for the square-based pyramid initial condition, appears to decrease with N_x for the MNIMC scheme, but its behaviour for the Upwind and Crank-Nicolson schemes is not as clear. More experiments are required to reveal the behaviour of the results when using these schemes. Therefore the results in Figure 5.8 are consistent with those of Figure 5.7, for the regularity initial conditions we predicted the initial conditions in Figure 5.8 might possess.

These are a limited set of results, so cannot be used to determine a definition for regularity of multiplicatively non-separable $u_0(x, y)$. The function $u_0(x, y)$ possessing a jump discontinuity will obviously play a role in the definition for the equivalent of $r_1 = 0$ and/or $r_2 = 0$. The results for the square-square ($r_1 = 0$ and $r_2 = 0$) and square-triangle ($r_1 = 0$ and $r_2 = 1$) could be used to imply that perhaps the regularity is determined by the minimum of r_1 and r_2, as both exhibit the same behaviour despite r_2 being different. The results for the 2D square and tent initial conditions also help to reinforce this idea along with the results from the analysis of the order of convergence of R_1 with respect to N_x. However this does not help with the definition of regularity for continuous functions and we should not let it determine a definition for regularity. Regularity should be defined through the proof of a bound on the continuous 2D Fourier coefficients of a multiplicatively non-separable function as was done for the 1D case.

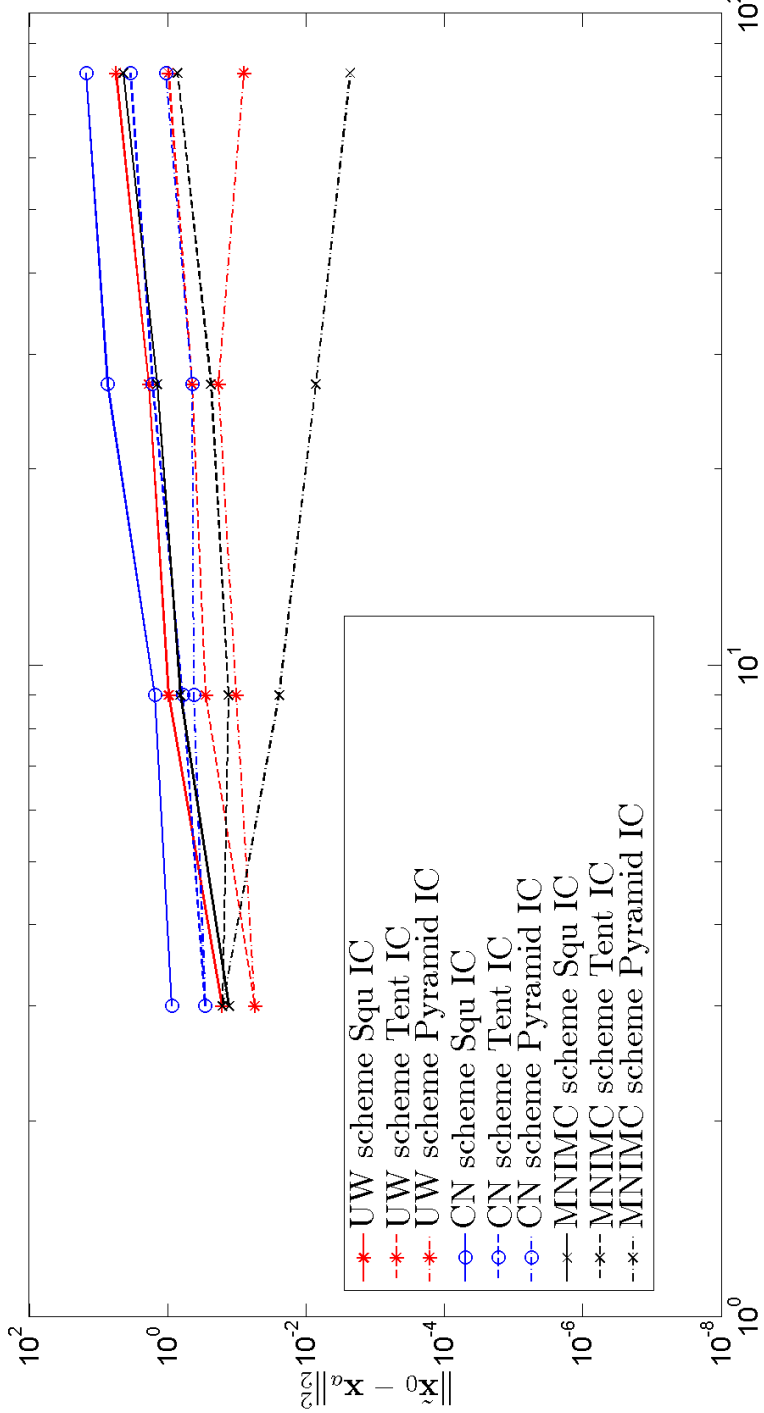


Figure 5.8: *The square of the l_2-norm of the error in the analysis vector, calculated through strong constraint 4D-Var data assimilation numerical experiments, under the influence of errors introduced by finite difference approximations in the forward model. The results were generated using the Upwind (UW), Crank-Nicolson (CN) and MNIMC schemes as the forward models for solving the 2D linear advection problem in (5.1), using $h_1 = h_2 = 0.5$, $\mu_1 = \mu_2 = 1$ and $L = 4$, where $N_x = N_y = 3^\alpha$ for $\alpha = 1, \ldots, 4$ ($\Delta t = \frac{1}{2N_x}$). The functions considered for $u_0(x, y)$ in these experiments are defined in Section 5.12 by multiplicatively non-separable 2D square (Squ IC), tent (Tent IC) and square-based pyramid (Pyramid IC) functions. The results are plotted using logarithmic scales to demonstrate the order of convergence as both N_x and N_y are increased at the same rate.*

5.13 Summary

This chapter has focused on extending the results of Chapters 3 and 4 to the 2D linear advection problem. The chapter began by setting out the 2D linear advection to be considered and outlining the relevant properties of the 2D Fourier series and 2D DFT. The definitions for numerical dissipation and dispersion outlined in Chapter 3 are sufficient for defining these terms for the 2D linear advection problem. Using these definitions, we chose to investigate the effects of numerical model error on the l_2-norm of the error in the analysis vector, through the Upwind and Crank-Nicolson schemes for the problem. The numerically dissipative and dispersive properties of these schemes is determined by the values of h_1 and h_2, the CFL numbers in the x- and y-directions. The values these variables can take is restricted by the numerical stability of the schemes.

The Fourier series method developed in Chapter 3 was used to construct the MNIMC for the 2D linear advection problem, a numerically non-dissipative and non-dispersive scheme, with respect to all resolvable wavenumber components of the numerical solution. Examining the aliasing error present in this scheme, we find it also has a shifted periodic nature in the x- and y-directions. The period of this shift is determined by the denominator of h_1 and h_2 respectively. Consequently, this scheme could be used to generate perfect observations of the physical system, both numerically and algebraically. However, we again make use of MATLABs *circshift* function [74] to generate perfect observations numerically. Using the MNIMC to construct perfect observations algebraically, we are able to construct the analysis vector similarly to Chapter 3. Continuing the work of the 1D linear advection problem into the 2D linear advection problem, has shown how the results can easily be extended to the d-dimensional linear advection problem, $d \in \mathbb{N}$.

In order to develop a bound for the l_2-norm of the error in the analysis vector using the spectral approach of Chapter 4, bounds were derived for the 2D Fourier coefficients and the error in the coefficients found through the 2D DFT, compared to the 2D Fourier coefficients for the same resolvable wavenumber component. These were defined for multiplicatively separable two-dimensional functions due to the problems associated with non-separable functions. However, the results for both types of function should be consistent. Deriving the bound for multiplicatively non-separable functions is a future extension of this work.

The bound on the error in the l_2-norm of the error in the analysis vector is important as it may be possible to use it to characterise the behaviour of the error with respect to the number of discretisation points when considering full sets of observations, the number of sets of observations in the assimilation window, the numerically dissipative and dispersive properties of the schemes and the smoothness of the true initial condition. This bound has many more summations which need to be analysed to find the dominant summation that determines the behaviour of the bound. Candidate summations for the dominant summation were identified based on the experience

gained from the 1D linear advection problem. Analysing these summations revealed that summation R_1 was dominant over the other candidate summations. The results from analysing the order of convergence of R_1 to zero with respect to N_x were compared against the same for the l_2-norm of the error in the analysis vector, obtained through strong constraint 4D-Var data assimilation numerical experiments. Both sets of data appeared to show that as the number of discretisation points is increased in the x- and y-directions simultaneously when considering full sets of observations, the error in the analysis vector increases for initial conditions containing lines of discontinuity. It is important to understand the reason for this increase as it is an unexpected result. Gibb's phenomenon in two-dimensions may be contributing to this error in some way. Larger values of N_x need to be considered in the strong constraint 4D-Var data assimilation numerical experiments before any conclusions on the behaviour of the error for other regularity initial conditions, or the ability of R_1 to represent the behaviour of the error, can be made.

Once this analysis has been completed, it would be interesting to re-introduce observation errors to the problem, to understand how the considered numerical model error and observation errors behave in unison for this physical system with respect to N_x. Finding an optimal number of discretisation points when considering full sets of observations, to perform strong constraint 4D-Var data assimilation, that minimises the effects of these errors would be of great use as the computational resources required to solve the problem could be justified. This analysis is left as future work.

CHAPTER 6

The 2D Linearised Shallow Water Problem

The linearised shallow water equations together with circulant boundary conditions and initial conditions, are considered as our physical system of interest for continuing our investigation into the effects of numerical model error on strong constraint 4D-Var data assimilation. The linearised shallow water equations are a logical next step for our investigation and present another meteorologically relevant physical system to consider. The shallow water equations form a coupled non-linear system which when linearised, under certain conditions, can decouple to form a system of linear advection equations. We will consider the 2D linearised shallow water equations so that a linear system of equations can be investigated, making use of the knowledge we have gained on 2D problems in Chapter 5. However we will choose assumptions for deriving our system of equations that ensure the system is linear and the equations do not decouple. If we do not make the latter assumption, we will not be able to learn much more than we have already learnt in Chapter 5 about investigating systems of equations.

The Chapter begins by deriving the 2D linearised shallow water equations from the non-linear shallow water equations. An analytical solution for the physical system is then derived, following the work of Cullen [87]. Finite difference schemes for solving the 2D linearised shallow water problem are then defined, highlighting the challenges that exist in choosing a basis which simultaneously diagonalises all considered finite difference schemes. We also attempt to define the equivalent of numerical dissipation and dispersion for a multivariate system of equations. Such a definition would allow the numerical model error introduced by these schemes to be categorised. The polar decomposition of matrices appears to be a useful tool towards this end and is discussed.

After having accessed the properties of the considered finite difference schemes, we turn our attention to the strong constraint 4D-Var data assimilation problem defined in Section 2.3. We begin our analysis similarly to Chapters 3 and 5, by considering the problem in the absence of all forms of error, other than those introduced by the finite difference schemes used as the forward model. This requires perfect observations

of the system. The MNIMC scheme is defined in the hope that it can be used to create perfect observations both algebraically and numerically. However, in this instance, we find we are unable to identify a method for generating perfect observations numerically, as the MNIMC scheme does not possess a shifted periodic nature for this problem. We end our analysis by discussing how this problem could be continued in the future.

6.1 The shallow water equations

In order define the shallow water equations (SWEs), we define the functions $u, v, h : \mathbb{R} \times \mathbb{R} \times [0, \infty) \to \mathbb{R}$, such that $(x, y, t) \mapsto u(x, y, t), v(x, y, t), h(x, y, t)$, and the function $b : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ such that $(x, y) \mapsto b(x, y)$. Then the two dimensional shallow water equations in conservative form are given by [69],

$$\frac{\partial}{\partial t}(uh) + \frac{\partial}{\partial x}(u^2 h) + \frac{\partial}{\partial y}(uvh) - fvh + gh\frac{\partial}{\partial x}(b + h) = 0, \tag{6.1}$$

$$\frac{\partial}{\partial t}(vh) + \frac{\partial}{\partial x}(uvh) + \frac{\partial}{\partial y}(v^2 h) + fuh + gh\frac{\partial}{\partial y}(b + h) = 0, \tag{6.2}$$

$$\frac{\partial h}{\partial t} + \frac{\partial}{\partial x}(hu) + \frac{\partial}{\partial y}(hv) = 0. \tag{6.3}$$

This is a non-linear hyperbolic system in two dimensions [6]. The functions u, v and h denote the *speed of the fluid in the x-direction*, *speed of the fluid in the y-direction* and the *height of the fluid flow above the bed of the fluid channel*, respectively. The height of the channel bed is given by $b(x, y)$, a function constant with respect to time. This is a simplification as in the real world, silt in the channel will move with the fluid flow. Modelling $b(x, y)$ accurately in a numerical model can be a challenge [88]. Source term splitting methods [89] and relaxation schemes [88] are examples of methods that have been used to attempt to accurately model source terms such as $b(x, y)$. The variable g denotes the *acceleration due to gravity* $g = 9.81\text{ms}^{-2}$, whilst $f \in \mathbb{R}$ denotes the *Coriolis acceleration*. Coriolis acceleration demonstrates the effect of the Earth's rotation on fluid motion and becomes a dominant variable in waves with large wavelengths [69]. Coriolis acceleration varies with the latitude of the Earth, but here we will assume f to be a fixed constant. This is known as the "mid-latitude f-plane assumption" [20]. These equations are derived from the Navier Stokes as shown by [9, 69]. Equations (6.1)-(6.3) are in fact the *depth averaged shallow water equations*, but are generally known as the shallow water equations [9]. Considering the depth averaged equations reveals the large scale behaviour of the considered system, neglecting small scale variations. In applications such as the modelling of tidal flows, floods and tsunami motion, where the height of the waves is small in comparison to the wavelength of the waves, it is the large scale motion of these waves that we are interested in [9].

We choose to examine the system in conservative form, as finite difference schemes solving non-conservative forms of equations can experience convergence problems when

considering shock profiles [72]. Shock profiles are of great interest when considering the shallow water equations. Dam break problems and tsunami waves both consist of waves with shock profiles [9]. Modelling these processes can be accomplished through solving the linearised shallow water equations due to their waves possessing a long wavelength when compared to their depth [9, 69]. Therefore, it is important that we choose our equations to facilitate such initial conditions. In the next Section we linearise the equations in (6.1)-(6.3) under the assumption $b(x, y) = 0$.

6.1.1 Linearising the shallow water equations

When linearising the shallow water equations in (6.1)-(6.3), there are two methods to choose from; freezing coefficients or perturbing the variables about a known solution [9]. We choose the latter option and perturb our functions u, v and h about a known solution, a steady uniform flow [69]. Let $(\overline{u}, \overline{v}, \overline{h})$ denote the steady uniform flow $\overline{u}, \overline{v}, \overline{h} \in \mathbb{R}$ and the functions $u', v', h' : \mathbb{R} \times \mathbb{R} \times [0, \infty) \to \mathbb{R}$, such that $(x, y, t) \mapsto u'(x, y, t), v'(x, y, t), h'(x, y, t)$ denote the perturbation from the steady state solution. The variables with $\overline{\cdot}$ must not be confused with the complex conjugate, especially since these are real-valued functions. We then substitute,

$$u = \overline{u} + u', \quad v = \overline{v} + v', \quad h = \overline{h} + h', \tag{6.4}$$

into equations (6.1)-(6.3). By neglecting the 2nd order terms and using the knowledge that the steady uniform flow satisfies equations (6.1)-(6.3), we achieve the linearised shallow water equations [69],

$$\frac{\partial u'}{\partial t} + u'\frac{\partial \overline{u}}{\partial x} + \overline{u}\frac{\partial u'}{\partial x} + v'\frac{\partial \overline{u}}{\partial y} + \overline{v}\frac{\partial u'}{\partial y} - fv' + g\frac{\partial h'}{\partial x} = 0, \tag{6.5}$$

$$\frac{\partial v'}{\partial t} + v'\frac{\partial \overline{u}}{\partial x} + \overline{u}\frac{\partial v'}{\partial x} + v'\frac{\partial \overline{v}}{\partial y} + \overline{v}\frac{\partial v'}{\partial y} + fu' + g\frac{\partial h'}{\partial y} = 0, \tag{6.6}$$

$$\frac{\partial h'}{\partial t} + \frac{\partial}{\partial x}(\overline{h}u' + h'\overline{u}) + \frac{\partial}{\partial y}(\overline{h}v' + h'\overline{v}) = 0. \tag{6.7}$$

Following the steady state solution choice of Cullen [87], we choose $(\overline{u}, \overline{v}, \overline{h}) = (0, 0, H)$, where $H \in \mathbb{R} \setminus \{0\}$ is a constant and substitute these into equations (6.5)-(6.7),

$$\frac{\partial u'}{\partial t} - fv' + g\frac{\partial h'}{\partial x} = 0, \tag{6.8}$$

$$\frac{\partial v'}{\partial t} + fu' + g\frac{\partial h'}{\partial y} = 0, \tag{6.9}$$

$$\frac{\partial h'}{\partial t} + H\frac{\partial u'}{\partial x} + H\frac{\partial v'}{\partial y} = 0. \tag{6.10}$$

These are the linearised shallow water equations in conservative form we will consider in our strong constraint 4D-Var problem [69].

Before defining any initial or boundary conditions for the linearised shallow water equations, we propose a change of variables. This change of variables will enable us to find an orthonormal basis for our analytical solution, in Section 6.1.2. Define the constant $\phi = gH$ known as the *geopotential* [39] and the function $\Phi' : \mathbb{R}\times\mathbb{R}\times[0,\infty) \to \mathbb{R}$ such that $(x, y, t) \mapsto \Phi'(x, y, t) = gh'(x, y, t)$. Our system of equations then becomes,

$$\frac{\partial u'}{\partial t} - fv' + \frac{\partial \Phi'}{\partial x} = 0, \tag{6.11}$$

$$\frac{\partial v'}{\partial t} + fu' + \frac{\partial \Phi'}{\partial y} = 0, \tag{6.12}$$

$$\frac{\partial \Phi'}{\partial t} + \phi\frac{\partial u'}{\partial x} + \phi\frac{\partial v'}{\partial y} = 0. \tag{6.13}$$

The quantity $\sqrt{\phi}$ is known as the *celerity* [9]. This change of variables is also used by Daley [20] and Lawless et al. [39].

We can re-write these equations in matrix-vector form by defining the vector functions $\mathbf{w}, \mathbf{w}' : \mathbb{R} \times \mathbb{R} \times [0,\infty) \to \mathbb{R}^3$ such that,

$$\begin{aligned}(x, y, t) \mapsto \mathbf{w}(x, y, t) &= [u'(x, y, t), v'(x, y, t), h'(x, y, t)]^T,\\ (x, y, t) \mapsto \mathbf{w}'(x, y, t) &= [u'(x, y, t), v'(x, y, t), \Phi'(x, y, t)]^T.\end{aligned}$$

Then equations (6.11)-(6.13), together with circulant boundary conditions and initial conditions becomes,

$$\frac{\partial \mathbf{w}'}{\partial t} + A\frac{\partial \mathbf{w}'}{\partial x} + B\frac{\partial \mathbf{w}'}{\partial y} + C\mathbf{w}' = \mathbf{0}, \qquad x, y \in [0,1)\times[0,1), \quad t > 0, \tag{6.14}$$

$$\begin{aligned}&\mathbf{w}'(x, y, t) = \mathbf{w}'(x + 1, y, t), && x, y \in \mathbb{R}\times\mathbb{R}, \quad t \geq 0,\\ &\mathbf{w}'(x, y, t) = \mathbf{w}'(x, y + 1, t), && x, y \in \mathbb{R}\times\mathbb{R}, \quad t \geq 0,\\ &\mathbf{w}'(x, y, 0) = [u_0(x, y), v_0(x, y), gh_0(x, y)]^T, && x, y \in \mathbb{R}\times\mathbb{R}\end{aligned} \tag{6.15}$$

where,

$$A = \begin{bmatrix} 0 & 0 & 1\\ 0 & 0 & 0\\ \phi & 0 & 0\end{bmatrix} \quad B = \begin{bmatrix} 0 & 0 & 0\\ 0 & 0 & 1\\ 0 & \phi & 0\end{bmatrix} \quad C = \begin{bmatrix} 0 & -f & 0\\ f & 0 & 0\\ 0 & 0 & 0\end{bmatrix}. \tag{6.16}$$

Here $\mathbf{0} \in \mathbb{R}^3$ denotes the zero vector. The initial conditions are defined by the functions $u_0, v_0, h_0 : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ such that $(x, y) \mapsto u_0(x, y), v_0(x, y), h_0(x, y)$. A consequence of applying circulant boundary conditions to (6.14)-(6.16) is that the functions u_0, v_0 and h_0 are one-periodic in both the x- and y-directions. The linearised shallow water equations are also considered in the context of data assimilation by Lawless et al. [39].

Solving (6.14)-(6.16) produces the solution (u', v', Φ'). We can transform the solution back into the variables (u', v', h') by transforming $\mathbf{w}'(x, y, t)$ using the invertible

matrix $\hat{G} \in \mathbb{R}^{3\times 3}$ such that $\mathbf{w} = \hat{G}^{-1}\mathbf{w}'$ where,

$$\hat{G} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & g \end{bmatrix}. \tag{6.17}$$

Now we have a linear system of equations in (6.14)-(6.16), which we can use as our physical system for our data assimilation problem. There is not a straight forward analytical solution to this problem, such as there was for the 1D and 2D linear advection equations in Chapters 3 and 5 respectively. In order to construct an analytical solution we must consider a Fourier series solution. As we wish to analyse the effects of numerical model error on strong constraint 4D-Var data assimilation in a similar way to Chapters 3 and 5, this formulation will help us. We derive the analytical solution in the following Section.

6.1.2 The Fourier series solution to the 2D linearised shallow water problem

The linearised shallow water equations in (6.14)-(6.16), are a linear system of equations with periodic boundary conditions, so an analytical solution can be found in the form of a Fourier series. As the system is in two dimensions in space, we can use the theory on 2D Fourier series set out in Section 5.2 of Chapter 5. The boundary conditions make the system one-periodic in both the x- and y-directions, hence the 2D Fourier series solution requires that $T_1 = T_2 = 1$.

The 2D Fourier series for the functions u', v' and Φ' are given by,

$$\begin{bmatrix} u'(x,y,t) \\ v'(x,y,t) \\ \Phi'(x,y,t) \end{bmatrix} \sim \sum_{p=-\infty}^{\infty}\sum_{q=-\infty}^{\infty} \begin{bmatrix} a_{p,q}(t) \\ b_{p,q}(t) \\ c_{p,q}(t) \end{bmatrix} e^{2\pi ipx} e^{2\pi iqy}, \tag{6.18}$$

where $a_{p,q}, b_{p,q}, c_{p,q} : [0,\infty) \to \mathbb{C}$, such that $t \mapsto a_{p,q}(t), b_{p,q}(t), c_{p,q}(t)$ for all $p, q \in \mathbb{Z}$. We then set,

$$\begin{bmatrix} a_{p,q}(t) \\ b_{p,q}(t) \\ c_{p,q}(t) \end{bmatrix} = \begin{bmatrix} i\hat{u}_{p,q} \\ i\hat{v}_{p,q} \\ \sqrt{\phi}\hat{\Phi}_{p,q} \end{bmatrix} e^{-2\pi i\omega_{p,q}t}, \tag{6.19}$$

where $\hat{u}_{p,q}, \hat{v}_{p,q}, \hat{\Phi}_{p,q} \in \mathbb{C}$ and $\omega_{p,q} \in \mathbb{R}$ are constants for all $p, q \in \mathbb{Z}$. Equation (6.19) is a modification of the functional dependence suggested by Daley [20, Equation (6.4.11), p. 195]. Setting $\hat{\mathbf{w}}_{p,q} = [\hat{u}_{p,q}, \hat{v}_{p,q}, \hat{\Phi}_{p,q}]^T$, the Fourier series for $\mathbf{w}'(x,y,t)$ is,

$$\mathbf{w}'(x,y,t) \sim \hat{C} \sum_{p=-\infty}^{\infty}\sum_{q=-\infty}^{\infty} \hat{\mathbf{w}}_{p,q} e^{2\pi i(px+qy-\omega_{p,q}t)}, \tag{6.20}$$

where,

$$\hat{C} = \begin{bmatrix} i & 0 & 0 \\ 0 & i & 0 \\ 0 & 0 & \sqrt{\phi} \end{bmatrix}. \tag{6.21}$$

As H is non-zero, $\hat{C}$ is invertible. A similar change of variables is suggested by Daley [20], when solving the linearised shallow water equations in terms of a stream-function, the velocity potential and the geopotential. We find that this change of variables is also useful in this problem, to create an orthonormal basis for the solution.

Equation (6.20) is a solution to the linearised shallow water equations in (6.14)-(6.16), when the Fourier series is convergent. Define, $\mathbf{S} : \mathbb{R} \times \mathbb{R} \times [0, \infty) \to \mathbb{R}^3$ such that,

$$(x, y, t) \mapsto \mathbf{S}(x, y, t) := \hat{C} \sum_{p=-\infty}^{\infty} \sum_{q=-\infty}^{\infty} \hat{\mathbf{w}}_{p,q} e^{2\pi i(px+qy-\omega_{p,q}t)}. \tag{6.22}$$

Our aim is to identify the constants $\omega_{p,q}$ and $\hat{\mathbf{w}}_{p,q}$, in $\mathbf{S}(x, y, t)$, to allow us to complete the Fourier series solution in (6.20). We use the method described by Cullen [87], to achieve this goal. Substituting (6.22) into (6.14) results in,

$$\hat{C} \sum_{p=-\infty}^{\infty} \sum_{q=-\infty}^{\infty} D_{p,q} \hat{\mathbf{w}}_{p,q} e^{2\pi i(px+qy-\omega_{p,q}t)} = \mathbf{0}, \tag{6.23}$$

where,

$$\begin{aligned} D_{p,q} &:= \hat{C}^{-1} \left\{ -2\pi i \omega_{p,q} I_3 + 2\pi i p A + 2\pi i q B + C \right\} \hat{C}, \\ &= \begin{bmatrix} -2\pi i \omega_{p,q} & -f & 2\pi p\sqrt{\phi} \\ f & -2\pi i \omega_{p,q} & 2\pi q\sqrt{\phi} \\ -2\pi p\sqrt{\phi} & -2\pi q\sqrt{\phi} & -2\pi i \omega_{p,q} \end{bmatrix}. \end{aligned} \tag{6.24}$$

The matrices A, B and C are defined in (6.16). In order to create $D_{p,q}$, we have post-multiplied by $I_3 = \hat{C}\hat{C}^{-1}$, where $I_3 \in \mathbb{R}^{3\times 3}$ is the 3×3 identity matrix. By including $\hat{C}^{-1}$ in the definition of $D_{p,q}$, it completes the change of variables implemented by the matrix $\hat{C}$. By the orthonormality of the Fourier basis functions, Equation (6.23) implies that,

$$\hat{C} D_{p,q} \hat{\mathbf{w}}_{p,q} = \mathbf{0}, \quad \text{for all } p, q \in \mathbb{Z}. \tag{6.25}$$

The non-zero solutions for $\hat{\mathbf{w}}_{p,q}$ are found when $\det(\hat{C}D_{p,q}) = 0$ [87]. As $\det(\hat{C})$ is non-zero, this is equivalent to $\det(D_{p,q}) = 0$. Define the matrix $\hat{D}_{p,q} \in \mathbb{R}^{3\times 3}$, such that $D_{p,q} = \hat{D}_{p,q} - 2\pi i \omega_{p,q} I_3$. Therefore,

$$\hat{D}_{p,q} = \begin{bmatrix} 0 & -f & 2\pi p\sqrt{\phi} \\ f & 0 & 2\pi q\sqrt{\phi} \\ -2\pi p\sqrt{\phi} & -2\pi q\sqrt{\phi} & 0 \end{bmatrix}. \tag{6.26}$$

Re-writing $D_{p,q}$ in this way allows us to see that when $\det(\hat{D}_{p,q}) = 0$, $2\pi i\omega_{p,q}$ is an eigenvalue of $\hat{D}_{p,q}$. Therefore, we can identify each $\omega_{p,q}$ through the eigenvalues of $\hat{D}_{p,q}$ [87]. By including $\hat{C}^{-1}$ in the definition of $D_{p,q}$ in (6.24), this ensured that $\hat{D}_{p,q}$ was a skew-symmetric matrix. As the eigenvalues of the matrix are also distinct, the eigenvectors of the matrix will be orthogonal for each $p, q \in \mathbb{Z}$ [90].

Define, $\hat{\omega}_{p,q} := \sqrt{\phi(p^2+q^2) + \left(\frac{f}{2\pi}\right)^2}$, then the eigenvalues of $\hat{D}_{p,q}$ are [87],

$$(\omega_1)_{p,q} = 0, \;\; (\omega_2)_{p,q} = \hat{\omega}_{p,q} \;\text{ and }\; (\omega_3)_{p,q} = -\hat{\omega}_{p,q}, \;\; \text{for all } p, q \in \mathbb{Z}. \tag{6.27}$$

Since $D_{p,q}(\hat{\mathbf{w}}_m)_{p,q} = \mathbf{0}$, we have that $\hat{D}_{p,q}(\hat{\mathbf{w}}_m)_{p,q} = 2\pi i(\omega_m)_{p,q}(\hat{\mathbf{w}}_m)_{p,q}$ for all $m = 1, 2, 3$ and $p, q \in \mathbb{Z}$. As a result, given any $(\omega_m)_{p,q}$ from (6.27), the corresponding eigenvector of $\hat{D}_{p,q}$ identifies each $(\hat{\mathbf{w}}_m)_{p,q}$ of the solution for all $m = 1, 2, 3$ and $p, q \in \mathbb{Z}$ [87]. The associated eigenvector of $(\omega_1)_{p,q} = 0$ is,

$$(\hat{\mathbf{w}}_1)_{p,q} = \frac{1}{2\pi\hat{\omega}_{p,q}} \begin{bmatrix} -2\pi q\sqrt{\phi} \\ 2\pi p\sqrt{\phi} \\ f \end{bmatrix} \;\text{ for all } p, q \in \mathbb{Z}. \tag{6.28}$$

This eigenvector has the property that $-(\hat{\mathbf{w}}_1)_{p,q} = (\hat{\mathbf{w}}_1)_{-p,-q}$ and is always a real eigenvector. The eigenvector associated with $(\omega_2)_{p,q} = \hat{\omega}_{p,q}$ is,

$$\begin{aligned} (\hat{\mathbf{w}}_2)_{p,q} &= \frac{1}{\sqrt{8}\pi\hat{\omega}_{p,q}\sqrt{p^2+q^2}} \begin{bmatrix} qf - 2\pi ip\hat{\omega}_{p,q} \\ -pf - 2\pi iq\hat{\omega}_{p,q} \\ 2\pi\sqrt{\phi}(p^2+q^2) \end{bmatrix} \;\text{ for all } p, q \in \mathbb{Z} \text{ not both zero and} \\ (\hat{\mathbf{w}}_2)_{0,0} &= \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -i \\ 0 \end{bmatrix}. \end{aligned} \tag{6.29}$$

The eigenvector associated with $\omega_3 = -\hat{\omega}_{p,q}$ is,

$$\begin{aligned} (\hat{\mathbf{w}}_3)_{p,q} &= \frac{1}{\sqrt{8}\pi\hat{\omega}_{p,q}\sqrt{p^2+q^2}} \begin{bmatrix} qf + 2\pi ip\hat{\omega}_{p,q} \\ -pf + 2\pi iq\hat{\omega}_{p,q} \\ 2\pi\sqrt{\phi}(p^2+q^2) \end{bmatrix} \;\text{ for all } p, q \in \mathbb{Z} \text{ not both zero and} \\ (\hat{\mathbf{w}}_3)_{0,0} &= \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ i \\ 0 \end{bmatrix}. \end{aligned} \tag{6.30}$$

The eigenvectors $(\hat{\mathbf{w}}_2)_{p,q}$ and $(\hat{\mathbf{w}}_3)_{p,q}$ have the property that $\overline{(\hat{\mathbf{w}}_2)_{p,q}} = (\hat{\mathbf{w}}_3)_{p,q}$ as their associated eigenvalues are complex conjugates.

Fixing m for some $m = 1, 2, 3$ and substituting the pair $\{(\omega_m)_{p,q}, (\hat{\mathbf{w}}_m)_{p,q}\}$ for each p, q into (6.22), creates a Fourier series solution $\mathbf{S}_m(x, y, t)$ to (6.14)-(6.16). When $m = 1$, the solution forms a *Rossby wave* [87]. In the case of the 2D linearised shallow

water equations, the Rossby wave remains constant over time, as $(\omega_1)_{p,q} = 0$ for all $p, q \in \mathbb{Z}$. The Fourier series solutions $\mathbf{S}_2(x, y, t)$ and $\mathbf{S}_3(x, y, t)$ are complex conjugate solutions, creating inertia-gravity waves [87]. Durran notes that "in many large-scale atmospheric and oceanic models, the Rossby waves are of greater physical significance than the faster-moving gravity waves" [91]. Therefore it is important that we model the Rossby wave solution correctly.

We now have three Fourier series solutions to the linearised shallow water problem in (6.14)-(6.16), of the form of (6.20). Then by the *principle of superposition*, the general Fourier series solution is,

$$\begin{aligned} &\mathbf{w}'_{p,q}(x, y, t) \\ \sim\ &\hat{C} \sum_{p=-\infty}^{\infty} \sum_{q=-\infty}^{\infty} \left\{ \alpha_{p,q}(\hat{\mathbf{w}}_1)_{p,q} + \beta_{p,q}(\hat{\mathbf{w}}_2)_{p,q} e^{-2\pi i \hat{\omega}_{p,q} t} + \gamma_{p,q}(\hat{\mathbf{w}}_3)_{p,q} e^{2\pi i \hat{\omega}_{p,q} t} \right\} e^{2\pi i (px+qy)}, \end{aligned} \tag{6.31}$$

where $\alpha_{p,q}, \beta_{p,q}, \gamma_{p,q} \in \mathbb{C}$ are constants determined by the initial condition of the problem. Transforming back into the original variables of the system, we achieve the solution,

$$\begin{aligned} &\mathbf{w}(x, y, t) \\ \sim\ &\hat{A} \sum_{p=-\infty}^{\infty} \sum_{q=-\infty}^{\infty} \left\{ \alpha_{p,q}(\hat{\mathbf{w}}_1)_{p,q} + \beta_{p,q}(\hat{\mathbf{w}}_2)_{p,q} e^{-2\pi i \hat{\omega}_{p,q} t} + \gamma_{p,q}(\hat{\mathbf{w}}_3)_{p,q} e^{2\pi i \hat{\omega}_{p,q} t} \right\} e^{2\pi i (px+qy)}, \end{aligned} \tag{6.32}$$

where $\hat{A} := \hat{G}^{-1}\hat{C}$ is an invertible matrix.

Now we consider how to identify the coefficients $\alpha_{p,q}$, $\beta_{p,q}$ and $\gamma_{p,q}$. Evaluating (6.32) at time $t = 0$, we create,

$$\mathbf{w}(x, y, 0) \sim \hat{A} \sum_{p=-\infty}^{\infty} \sum_{q=-\infty}^{\infty} \left\{ \alpha_{p,q}(\hat{\mathbf{w}}_1)_{p,q} + \beta_{p,q}(\hat{\mathbf{w}}_2)_{p,q} + \gamma_{p,q}(\hat{\mathbf{w}}_3)_{p,q} \right\} e^{2\pi i (px+qy)}. \tag{6.33}$$

Notice that the eigenvectors of $\hat{D}_{p,q}$, given by $\{(\hat{\mathbf{w}}_m)_{p,q}\}_{m=1}^3$, form an orthonormal eigenbasis for $\mathbb{C}^3$. Suppose we also have a Fourier series for the initial condition $\mathbf{w}_0(x, y)$,

$$\mathbf{w}_0(x, y) \sim \sum_{p=-\infty}^{\infty} \sum_{q=-\infty}^{\infty} \mathbf{c}_{p,q} e^{2\pi i (px+qy)}, \tag{6.34}$$

where $\mathbf{c}_{p,q} \in \mathbb{C}^3$ are constants defined by,

$$\mathbf{c}_{p,q} = \int_0^1 \int_0^1 \mathbf{w}_0(x, y) e^{-2\pi i (px+qy)} dx dy, \tag{6.35}$$

for all $p, q \in \mathbb{Z}$. Then,

$$\mathbf{c}_{p,q} = \hat{A}[\alpha_{p,q}(\hat{\mathbf{w}}_1)_{p,q} + \beta_{p,q}(\hat{\mathbf{w}}_2)_{p,q} + \gamma_{p,q}(\hat{\mathbf{w}}_3)_{p,q}]. \tag{6.36}$$

As the eigenvectors of $\hat{D}_{p,q}$ form an orthonormal eigenbasis for $\mathbb{C}^3$, $\hat{A}^{-1}\mathbf{c}_{p,q}$ can be constructed completely from this basis. Let $E_{p,q} \in \mathbb{C}^{3\times 3}$ be such that the mth column of $E_{p,q}$ is $(\hat{\mathbf{w}}_m)_{p,q}$ for all $p, q \in \mathbb{Z}$. Then (6.36) can be re-written as [87],

$$\hat{A}^{-1}\mathbf{c}_{p,q} = E_{p,q}[\alpha_{p,q}, \beta_{p,q}, \gamma_{p,q}]^T. \tag{6.37}$$

The orthonormality of the eigenbasis results in $E_{p,q}^{-1} = E_{p,q}^*$ for all $p, q \in \mathbb{Z}$, where we remind the reader that $\cdot^*$ denotes the Hermitian of a matrix. The coefficients $\alpha_{p,q}$, $\beta_{p,q}$ and $\gamma_{p,q}$ can then be found by applying $E_{p,q}^*$ to $\hat{A}^{-1}\mathbf{c}_{p,q}$. Consequently,

$$\mathbf{w}(x, y, 0) \sim \sum_{p=-\infty}^{\infty} \sum_{q=-\infty}^{\infty} \hat{A}E_{p,q}(E_{p,q}^*\hat{A}^{-1}\mathbf{c}_{p,q})e^{2\pi i(px+qy)}. \tag{6.38}$$

In order to progress this system forward t in time, $\alpha_{p,q}$, $\beta_{p,q}$ and $\gamma_{p,q}$ need to be multiplied by 1, $e^{-2\pi i\hat{\omega}_{p,q}}$ and $e^{2\pi i\hat{\omega}_{p,q}}$ respectively. Therefore,

$$\mathbf{w}(x, y, t) \sim \sum_{p=-\infty}^{\infty} \sum_{q=-\infty}^{\infty} \hat{A}E_{p,q}Q_{p,q}^t E_{p,q}^*\hat{A}^{-1}\mathbf{c}_{p,q}e^{2\pi i(px+qy)}, \tag{6.39}$$

where,

$$Q_{p,q} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & e^{-2\pi i\hat{\omega}_{p,q}} & 0 \\ 0 & 0 & e^{2\pi i\hat{\omega}_{p,q}} \end{bmatrix}. \tag{6.40}$$

It is useful to notice here that $Q_{p,q}$ contains the exponential of the eigenvalues of $\hat{D}_{p,q}$ along its main diagonal, ie: $Q_{p,q} = e^{\Gamma_{p,q}}$ where,

$$\Gamma_{p,q} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -2\pi i\hat{\omega}_{p,q} & 0 \\ 0 & 0 & 2\pi i\hat{\omega}_{p,q} \end{bmatrix}. \tag{6.41}$$

Therefore,

$$E_{p,q}Q_{p,q}^t E_{p,q}^* = E_{p,q}e^{\Gamma_{p,q}t}E_{p,q}^* = e^{E_{p,q}\Gamma_{p,q}E_{p,q}^*t} = e^{\hat{D}_{p,q}t}, \tag{6.42}$$

for all $p, q \in \mathbb{Z}$ and

$$\mathbf{w}(x, y, t) \sim \sum_{p=-\infty}^{\infty} \sum_{q=-\infty}^{\infty} \hat{A}e^{\hat{D}_{p,q}t}\hat{A}^{-1}\mathbf{c}_{p,q}e^{2\pi i(px+qy)}. \tag{6.43}$$

This is our Fourier series solution to the 2D linearised shallow water problem in (6.14)-(6.16).

Examining this solution, we can see that the matrix $e^{\hat{D}_{p,q}t}$, implements the evolution of the solution over time. The matrix $\hat{D}_{p,q}$ contains the elements of the vector $2\pi\hat{\omega}_{p,q}(\hat{\mathbf{w}}_1)_{p,q}$ arranged in such a way that for any vector $\mathbf{z} \in \mathbb{C}^3$,

$$\hat{D}_{p,q}\mathbf{z} = 2\pi\hat{\omega}_{p,q}(\hat{\mathbf{w}}_1)_{p,q} \wedge \mathbf{z}, \tag{6.44}$$

where $\wedge$ denotes the vector cross-product. Then post-multiplying $\hat{A}^{-1}\mathbf{c}_{p,q}$ by $e^{\hat{D}_{p,q}t}$ in (6.43), applies a rotation to the vector $\hat{A}^{-1}\mathbf{c}_{p,q}$ around an axis in the direction $2\pi\hat{\omega}_{p,q}(\hat{\mathbf{w}}_1)_{p,q}$, an angle t [92]. Due to the form of $\hat{\omega}_{p,q}$, it is not possible to pick a time $t > 0$ such that the system can be rotated enough to recover the initial conditions.

In Section 5.2 of Chapter 5, we discussed that the coefficients of the (p, q)th and $(-p, -q)$th wavenumber components of a 2D Fourier series are complex conjugates. Consider the complex conjugate of $\mathbf{c}_{p,q}$ in (6.36) and compare it against $\mathbf{c}_{-p,-q}$ in the form of (6.36). Using the knowledge that $\overline{\hat{A}(\hat{\mathbf{w}}_1)_{p,q}} = \hat{A}(\hat{\mathbf{w}}_1)_{-p,-q}$, $\overline{\hat{A}(\hat{\mathbf{w}}_2)_{p,q}} = \hat{A}(\hat{\mathbf{w}}_2)_{-p,-q}$ and $\overline{\hat{A}(\hat{\mathbf{w}}_3)_{p,q}} = \hat{A}(\hat{\mathbf{w}}_3)_{-p,-q}$, we find that,

$$\alpha_{-p,-q} = \overline{\alpha_{p,q}}, \quad \beta_{-p,-q} = \overline{\beta_{p,q}} \quad \text{and} \quad \gamma_{-p,-q} = \overline{\gamma_{p,q}}. \tag{6.45}$$

When considering the 1D linearised shallow water equations in the absence of Coriolis acceleration, ie: $f = 0$, we find that the system of equations decomposes into a system of 1D linear advection equations [9]. If we consider the 2D linearised shallow water problem in (6.14)-(6.16) when $f = 0$, we have that $C = 0 \in \mathbb{R}^{3\times3}$, the 3×3 matrix of zeros. Therefore we only need to consider the matrices A and B to determine if the system can be decoupled. As the matrices A and B do not commute, the matrices are not simultaneously diagonalisable [93], so the system of equations cannot be decoupled into a system of 2D linear advection equations.

In the next Section, we define two finite difference schemes which can be used to solve the 2D linearised shallow water problem in (6.14)-(6.16). These finite difference schemes will form the forward model in our strong constraint 4D-Var data assimilation problem, defined in Section 2.3, when considering the 2D linearised shallow water problem as our physical system.

6.2 Finite difference schemes for solving the 2D linearised shallow water problem

As with the 1D and 2D linear advection problems, we need a way to formulate our finite difference schemes for solving the 2D linearised shallow water problem in (6.14)-(6.16). We will use the structure of our schemes to guide their implementation. There are many finite difference schemes we could choose to solve the 2D linearised shallow water problem. As with the 2D linear advection problem, we will use the Upwind and Crank-Nicolson schemes to investigate the effects of numerical model error on the results of

strong constraint 4D-Var data assimilation. This is due to their different numerically dissipative and dispersive properties and their easy implementation for the 2D linear advection problem. Initially we will use this justification as we will see in Section 6.3, that defining numerical dissipation and dispersion for a multivariate system is not an easy process.

Define the equally spaced mesh over $[0,1] \times [0,1]$, with spatial steps $\Delta x = \frac{1}{N_x}$ and $\Delta y = \frac{1}{N_y}$ together with the time step $\Delta t \in \mathbb{R}^+$, as in Section 5.3 of Chapter 5. Let $U^n_{j,k}$, $V^n_{j,k}$, $\Phi^n_{j,k}$ and $H^n_{j,k}$ be the numerical solution at (x_j, y_k, t^n), approximating the solutions $u'(x_j, y_k, t^n)$, $v'(x_j, y_k, t^n)$, $\Phi'(x_j, y_k, t^n)$ and $h'(x_j, y_k, t^n)$, respectively for $j = 0, \ldots, N_x$, $k = 0, \ldots, N_y$ and $n \in \mathbb{N}_0$. We also define the vectors $\mathbf{U}^n, \mathbf{V}^n, \mathbf{\Phi}^n, \mathbf{H}^n \in \mathbb{R}^{N_x N_y}$ such that $\{\mathbf{U}^n\}_{(k-1)N_x+j} = U^n_{j-1,k-1}$, $\{\mathbf{V}^n\}_{(k-1)N_x+j} = V^n_{j-1,k-1}$, $\{\mathbf{\Phi}^n\}_{(k-1)N_x+j} = \Phi^n_{j-1,k-1}$ and $\{\mathbf{H}^n\}_{(k-1)N_x+j} = H^n_{j-1,k-1}$ for $j = 1, \ldots, N_x$ and $k = 1, \ldots, N_y$. We now have a regularly spaced grid over $[0,1] \times [0,1]$ to solve the 2D linearised shallow water problem and vectors to store the numerical solution.

Next we will consider our chosen schemes for solving the 2D linearised shallow water problem. The vector formulation of the problem in (6.14)-(6.16) allows us to easily construct the scheme so that it numerically solves for the variables (u', v', Φ'). We construct the vector $\hat{\mathbf{W}}^n_{j,k} \in \mathbb{R}^3$ for $j = 0, \ldots, N_x - 1$, $k = 0, \ldots, N_y - 1$ and $n \in \mathbb{N}_0$ such that,

$$\left\{\hat{\mathbf{W}}^n_{j,k}\right\}_q = \begin{cases} U^n_{j,k}, & \text{for } q = 1, \\ V^n_{j,k}, & \text{for } q = 2, \\ \Phi^n_{j,k}, & \text{for } q = 3, \end{cases} \tag{6.46}$$

to store the numerical solution for (u', v', Φ') at (x_j, y_k, t^n). However, our original problem was posed in the variables (u', v', h'), so we need to change the variables of our numerical solution. Define the vector $\mathbf{W}^n_{j,k} \in \mathbb{R}^3$ such that $\mathbf{W}^n_{j,k} = \hat{G}^{-1}\hat{\mathbf{W}}^n_{j,k}$. Then we have that,

$$\left\{\mathbf{W}^n_{j,k}\right\}_q = \begin{cases} U^n_{j,k}, & \text{for } q = 1, \\ V^n_{j,k}, & \text{for } q = 2, \\ H^n_{j,k}, & \text{for } q = 3. \end{cases} \tag{6.47}$$

The Upwind and Crank-Nicolson schemes are then defined to solve problem (6.14)-(6.16) numerically and apply the change of variables as follows:

- The Upwind scheme (explicit scheme) [59],

$$\begin{aligned} \mathbf{W}^{n+1}_{j,k} &= \hat{G}^{-1}\left[\left(I_3 + \frac{\Delta t}{\Delta x}A + \frac{\Delta t}{\Delta y}B - \Delta t C\right)\hat{G}\mathbf{W}^n_{j,k} - \frac{\Delta t}{\Delta x}A\hat{G}\mathbf{W}^n_{j-1,k}\right. \\ &\quad \left. - \frac{\Delta t}{\Delta y}B\hat{G}\mathbf{W}^n_{j,k-1}\right]. \end{aligned} \tag{6.48}$$

- The Crank-Nicolson scheme (implicit scheme) [70],

$$\begin{aligned}
&\hat{G}^{-1}\left[\left(I_3+\frac{\Delta t}{2}C\right)\hat{G}\mathbf{W}_{j,k}^{n+1}+\frac{\Delta t}{2\Delta x}A\hat{G}\left(\mathbf{W}_{j+1,k}^{n+1}-\mathbf{W}_{j-1,k}^{n+1}\right)\right.\\
&\left.+\frac{\Delta t}{2\Delta y}B\hat{G}\left(\mathbf{W}_{j,k+1}^{n+1}-\mathbf{W}_{j,k-1}^{n+1}\right)\right]\\
=\;&\hat{G}^{-1}\left[\left(I_3-\frac{\Delta t}{2}C\right)\hat{G}\mathbf{W}_{j,k}^{n}-\frac{\Delta t}{2\Delta x}A\hat{G}\left(\mathbf{W}_{j+1,k}^{n}-\mathbf{W}_{j-1,k}^{n}\right)\right.\\
&\left.-\frac{\Delta t}{2\Delta y}B\hat{G}\left(\mathbf{W}_{j,k+1}^{n}-\mathbf{W}_{j,k-1}^{n}\right)\right].
\end{aligned}\tag{6.49}$$

We can now implement these schemes in a similar way to their implementation for the 2D linear advection problem. In the 2D linear advection problem, we stacked the state of the system in a vector, as in (5.7). As we now solve for a vector describing the state of the system instead of a single variable, we stack the vectors producing the state of the system, in the same way. Define the vector $\mathbf{Q}_k^n \in \mathbb{R}^{3N_x}$ such that,

$$\mathbf{Q}_k^n = \left[(\mathbf{W}_{0,k}^n)^T, (\mathbf{W}_{1,k}^n)^T, \ldots, (\mathbf{W}_{N_x-1,k}^n)^T\right]^T, \tag{6.50}$$

for $k = 0, \ldots, N_y - 1$. We then stack the vectors $\mathbf{Q}_k^n$ to create $\mathbf{Z}^n \in \mathbb{R}^{3N_xN_y}$,

$$\mathbf{Z}^n = \left[(\mathbf{Q}_0^n)^T, (\mathbf{Q}_1^n)^T, \ldots, (\mathbf{Q}_{N_y-1}^n)^T\right]^T. \tag{6.51}$$

This vector contains the state of the numerical solution at time t^n and is structured such that,

$$[\mathbf{Z}^n]_{(k-1)3N_x+(j-1)3+s} = \begin{cases} U_{j-1,k-1}^n, & \text{for } s = 1, \\ V_{j-1,k-1}^n, & \text{for } s = 2, \\ H_{j-1,k-1}^n, & \text{for } s = 3, \end{cases} \tag{6.52}$$

for $j = 1, \ldots, N_x$ and $k = 1, \ldots, N_y$. This allows us to construct a matrix $M \in \mathbb{R}^{3N_xN_y \times 3N_xN_y}$, such that $\mathbf{Z}^{n+1} = M\mathbf{Z}^n$, for all $n \in \mathbb{N}_0$. This advances the numerical solution Δt through time and results in $N = 3N_xN_y$, where N was defined in Section 2.3. We will construct M for the Upwind scheme in order to demonstrate its block diagonal structure. The schematic for the Upwind scheme gives that,

$$\begin{aligned}
\mathbf{Q}_0^{n+1} &= R_1\mathbf{Q}_0^n + R_2\mathbf{Q}_{N_y-1}^n, \\
\mathbf{Q}_k^{n+1} &= R_1\mathbf{Q}_k^n + R_2\mathbf{Q}_{k-1}^n, \quad \text{for } k = 1, \ldots, N_y - 1,
\end{aligned}\tag{6.53}$$

where

$$R_1 = \begin{bmatrix} R_{1a} & 0 & & \dots & R_{1b} \\ R_{1b} & R_{1a} & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & R_{1b} & R_{1a} & 0 \\ 0 & \dots & & R_{1b} & R_{1a} \end{bmatrix} \text{ and } R_2 = \begin{bmatrix} R_{2a} & 0 & & \dots & 0 \\ 0 & R_{2a} & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & R_{2a} & 0 \\ 0 & \dots & & 0 & R_{2a} \end{bmatrix}, \tag{6.54}$$

such that

- $R_{1a} = \hat{G}^{-1}\left(I_3 + \frac{\Delta t}{\Delta x}A + \frac{\Delta t}{\Delta y}B - \Delta t C\right)\hat{G}$,
- $R_{1b} = -\frac{\Delta t}{\Delta x}\hat{G}^{-1}A\hat{G}$,
- $R_{2a} = -\frac{\Delta t}{\Delta y}\hat{G}^{-1}B\hat{G}$.

The 0's within the matrices R_1 and R_2 are 3×3 zero matrices. We see that the matrices R_1 and R_2 are block circulant matrices, where the blocks are constructed from 3×3 matrices. Using the construction of $\mathbf{Z}^n$ in (6.51), we see that the matrix M is constructed as a block circulant matrix [65], where each block is of size $3N_x \times 3N_x$ such that,

$$M = \begin{bmatrix} R_1 & 0 & & \dots & R_2 \\ R_2 & R_1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & R_2 & R_1 & 0 \\ 0 & \dots & & R_2 & R_1 \end{bmatrix}. \tag{6.55}$$

Here the 0's denote $3N_x \times 3N_x$ zero matrices. The block circulant nature of M and its constituent blocks, is due to the circulant boundary conditions of the problem. The Crank-Nicolson scheme is constructed in a similar way.

We have seen that for the 1D and 2D linear advection problems, that the matrices implementing the chosen finite difference schemes, can all be diagonalised using the relevant DFT. In the following Section, we construct a matrix which when applied to M, will apply the 2D DFT. Unlike the 1D and 2D linear advection problems, as the 2D linearised shallow water problem is a system of PDEs, this matrix will not diagonalise the matrix M. However, it will allow us to examine the amplification matrices which construct M and propagate the numerical solution forward in time, similarly to the eigenvalues in the previous chapters.

6.2.1 The 2D discrete Fourier transform

In Section 5.3.1 we defined the matrix $V \in \mathbb{C}^{N_xN_y \times N_xN_y}$ in Equation (5.16), such that by applying V^* to any vector $\mathbf{z} \in \mathbb{R}^{N_xN_y}$, we apply the 2D DFT to $\mathbf{z}$. This identifies the coefficients of the 2D discrete Fourier series for $\mathbf{z}$, in vector form. In the

case of the 2D linearised shallow water problem, we can construct a similar matrix $X \in \mathbb{C}^{3N_xN_y \times 3N_xN_y}$, using the entries of V. The result is that $X^*\mathbf{Z}^n$ identifies the coefficients for the 2D discrete Fourier series of each $\mathbf{W}^n_{j,k}$, making up $\mathbf{Z}^n$.

The matrix X is constructed from block diagonal matrices. Denote the (p,q)th block constructing X by $X_{p,q} \in \mathbb{C}^{3\times 3}$. Then X is constructed such that,

$$X = \begin{bmatrix} X_{1,1} & X_{1,2} & \dots & X_{1,N_xN_y} \\ X_{2,1} & X_{2,2} & \dots & X_{2,N_xN_y} \\ \vdots & \vdots & \vdots & \vdots \\ X_{N_xN_y,1} & X_{N_xN_y,2} & \dots & X_{N_xN_y,N_xN_y} \end{bmatrix}., \tag{6.56}$$

where the blocks are defined by,

$$X_{p,q} = \{V\}_{p,q} I_3, \quad \text{for } p,q = 1,\dots,N_xN_y. \tag{6.57}$$

This is the (p,q)th element of V, multiplied by the 3×3 identity matrix. This leads the elements of the matrix X to be defined by,

$$\begin{aligned} & \{X\}_{(k-1)3N_x+(j-1)3+r,(q-1)3N_x+(p-1)3+s} \\ = \; & \left\{X_{(k-1)N_x+j,(q-1)N_x+p}\right\}_{r,s} \\ = \; & \frac{1}{\sqrt{N_xN_y}} e^{\frac{2\pi i(p-1)(j-1)}{N_x}} e^{\frac{2\pi i(q-1)(k-1)}{N_y}} \delta_{r,s}, \end{aligned} \tag{6.58}$$

for $j = 1,\dots,N_x$, $k = 1,\dots,N_y$ and $r = 1,2,3$. This is a symmetric matrix by the symmetry of its structure and the matrix V.

Suppose we now apply X^* to $\mathbf{Z}^n$,

$$
\begin{aligned}
& \{X^*\mathbf{Z}^n\}_{(k-1)3N_x+(j-1)3+r} \\
= & \sum_{p=1}^{3N_xN_y} \{X^*\}_{(k-1)3N_x+(j-1)3+r,p}\{\mathbf{Z}^n\}_p, \\
= & \sum_{q=1}^{N_y}\sum_{p=1}^{N_x}\sum_{s=1}^{3}\{\overline{X}\}_{(k-1)3N_x+(j-1)+r,(q-1)3N_x+(p-1)3+s}\{\mathbf{Z}^n\}_{(q-1)3N_x+(p-1)3+s}, \\
& \text{by the symmetry of } X, \\
= & \sum_{q=1}^{N_y}\sum_{p=1}^{N_x}\sum_{s=1}^{3}\frac{1}{\sqrt{N_xN_y}}e^{\frac{-2\pi i(p-1)(j-1)}{N_x}}e^{\frac{-2\pi i(q-1)(k-1)}{N_y}}\delta_{r,s}\{\mathbf{Z}^n\}_{(q-1)3N_x+(p-1)3+s}, \\
= & \sum_{q=1}^{N_y}\sum_{p=1}^{N_x}\frac{1}{\sqrt{N_xN_y}}e^{\frac{-2\pi i(p-1)(j-1)}{N_x}}e^{\frac{-2\pi i(q-1)(k-1)}{N_y}}\{\mathbf{Z}^n\}_{(q-1)3N_x+(p-1)3+r}, \\
= & \begin{cases} \frac{1}{\sqrt{N_xN_y}}\sum_{p=1}^{N_x}\sum_{q=1}^{N_y}U^n_{p-1,q-1}e^{\frac{-2\pi i(p-1)(j-1)}{N_x}}e^{\frac{-2\pi i(q-1)(k-1)}{N_y}}, & \text{for } r=1, \\ \frac{1}{\sqrt{N_xN_y}}\sum_{p=1}^{N_x}\sum_{q=1}^{N_y}V^n_{p-1,q-1}e^{\frac{-2\pi i(p-1)(j-1)}{N_x}}e^{\frac{-2\pi i(q-1)(k-1)}{N_y}}, & \text{for } r=2, \\ \frac{1}{\sqrt{N_xN_y}}\sum_{p=1}^{N_x}\sum_{q=1}^{N_y}H^n_{p-1,q-1}e^{\frac{-2\pi i(p-1)(j-1)}{N_x}}e^{\frac{-2\pi i(q-1)(k-1)}{N_y}}, & \text{for } r=3, \end{cases}
\end{aligned} \tag{6.59}
$$

for $j = 1, \ldots, N_x$ and $k = 1, \ldots, N_y$. Depending on the value of r, we acquire the coefficients required to create a 2D discrete Fourier series for $\mathbf{U}^n$, $\mathbf{V}^n$ and $\mathbf{H}^n$. This gives $X^*\mathbf{Z}^n$ the following structure, where $\mathcal{F}_{p,q}(\cdot)$ for $p = 1, \ldots, N_x$ and $q = 1, \ldots, N_y$ is defined in Section 5.3.1,

$$
X^*\mathbf{Z}^n = \begin{bmatrix} \mathcal{F}_{1,1}(\mathbf{U}^n) \\ \mathcal{F}_{1,1}(\mathbf{V}^n) \\ \mathcal{F}_{1,1}(\mathbf{H}^n) \\ \vdots \\ \mathcal{F}_{N_x,1}(\mathbf{U}^n) \\ \mathcal{F}_{N_x,1}(\mathbf{V}^n) \\ \mathcal{F}_{N_x,1}(\mathbf{H}^n) \\ \vdots \\ \mathcal{F}_{1,N_y}(\mathbf{U}^n) \\ \mathcal{F}_{1,N_y}(\mathbf{V}^n) \\ \mathcal{F}_{1,N_y}(\mathbf{H}^n) \\ \vdots \\ \mathcal{F}_{N_x,N_y}(\mathbf{U}^n) \\ \mathcal{F}_{N_x,N_y}(\mathbf{V}^n) \\ \mathcal{F}_{N_x,N_y}(\mathbf{H}^n) \end{bmatrix}. \tag{6.60}
$$

As the numerical solution is equal to the initial condition at each grid point when $n = 0$,

(6.60) creates the coefficients for the 2D discrete Fourier series of $u'(x,y,0)$, $v'(x,y,0)$ and $h'(x,y,0)$, when the Fourier series is convergent.

The matrix $K = X^*MX$, $K \in \mathbb{C}^{3N_xN_y \times 3N_xN_y}$, has a block diagonal structure for the Upwind and Crank-Nicolson schemes. The blocks constructing K are the matrices $K_{p,q} \in \mathbb{C}^{3\times 3}$ for $p = 1,\ldots,N_x$ and $q = 1,\ldots,N_y$, which run along the main diagonal of K,

$$K = \begin{bmatrix} K_{1,1} & 0 & \ldots & 0 & \ldots & \ldots & \ldots & 0 \\ 0 & K_{2,1} & \ldots & 0 & \ldots & \ldots & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & & & & \vdots \\ 0 & 0 & \ldots & K_{N_x,1} & \ldots & \ldots & \ldots & 0 \\ \vdots & \vdots & & & \ddots & & & \vdots \\ 0 & 0 & \ldots & \ldots & \ldots & K_{1,N_y} & \ldots & 0 \\ \vdots & \vdots & & & & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \ldots & \ldots & 0 & \ldots & K_{N_x,N_y} \end{bmatrix}. \tag{6.61}$$

In equation form this gives,

$$\{K\}_{(k-1)3N_x+(j-1)3+r,(q-1)3N_x+(p-1)3+s} = \{K_{p,q}\}_{r,s}\delta_{k,q}\delta_{j,p}, \tag{6.62}$$

for $p = 1,\ldots,N_x$ and $q = 1,\ldots,N_y$. We then have that $M = XKX^*$, so that,

$$\mathbf{Z}^1 = M\mathbf{Z}^0 = XKX^*\mathbf{Z}^0 \Rightarrow X^*\mathbf{Z}^1 = KX^*\mathbf{Z}^0. \tag{6.63}$$

Using equation (6.60), this reveals that,

$$\begin{bmatrix} \mathcal{F}_{p,q}(\mathbf{U}^1) \\ \mathcal{F}_{p,q}(\mathbf{V}^1) \\ \mathcal{F}_{p,q}(\mathbf{\Phi}^1) \end{bmatrix} = K_{p,q} \begin{bmatrix} \mathcal{F}_{p,q}(\mathbf{U}^0) \\ \mathcal{F}_{p,q}(\mathbf{V}^0) \\ \mathcal{F}_{p,q}(\mathbf{\Phi}^0) \end{bmatrix}. \tag{6.64}$$

So the matrix $K_{p,q}$ progresses the (p,q)th wavenumber component of the numerical solution, forward Δt in time. This leads us to referring to these matrices as *amplification matrices*. The amplification matrices for our considered finite difference schemes are as follows:

- The Upwind scheme,

$$K_{p,q} = \hat{A}\hat{C}^{-1}\left\{I_3 + \left(1 - e^{\frac{-2\pi i(p-1)}{N_x}}\right)\frac{\Delta t}{\Delta x}A + \left(1 - e^{\frac{-2\pi i(q-1)}{N_y}}\right)\frac{\Delta t}{\Delta y}B - \Delta t C\right\}\hat{C}\hat{A}^{-1}. \tag{6.65}$$

- The Crank Nicolson scheme,

$$\begin{aligned} & K_{p,q} \\ = \; & \hat{A}\hat{C}^{-1}\left\{I_3 + \frac{\Delta t}{2}C + i\frac{\Delta t}{\Delta x}\sin\left[\frac{2\pi(p-1)}{N}\right]A + i\frac{\Delta t}{\Delta y}\sin\left[\frac{2\pi(q-1)}{N_y}\right]B\right\}^{-1} \cdot \\ & \left\{I_3 - \frac{\Delta t}{2}C - i\frac{\Delta t}{\Delta x}\sin\left[\frac{2\pi(p-1)}{N}\right]A - i\frac{\Delta t}{\Delta y}\sin\left[\frac{2\pi(q-1)}{N_y}\right]B\right\}\hat{C}\hat{A}^{-1}. \end{aligned} \tag{6.66}$$

The numerical stability of these schemes is calculated via the eigenvalues of the matrix M implementing the scheme. The eigenvalues of M can be identified as follows,

$$\begin{aligned} & \det\left(M - \lambda I_{3N_xN_y}\right) = 0, \\ \Rightarrow \; & \det\left(K - \lambda I_{3N_xN_y}\right) = 0, \\ \Rightarrow \; & \prod_{p=1}^{N_x}\prod_{q=1}^{N_y}\det\left(K_{p,q} - \lambda I_3\right) = 0, \end{aligned} \tag{6.67}$$

due to the block diagonal structure of K [94]. Hence it is the eigenvalues of the amplification matrices which determine the numerical stability of the schemes.

Diagonalising the matrices $K_{p,q}$ results in $K_{p,q} = V_{p,q}\Lambda_{p,q}V_{p,q}^{-1}$, where $V_{p,q}, \Lambda_{p,q} \in \mathbb{C}^{3\times 3}$ for all $p = 1, \ldots, N_x$ and $q = 1, \ldots, N_y$. The matrix $\Lambda_{p,q}$ is a diagonal matrix with the eigenvalues of $K_{p,q}$ along its main diagonal. The matrix $V_{p,q}$ contains the eigenvectors of $K_{p,q}$ as its columns, in the same order as the corresponding eigenvalues appear along the main diagonal of $\Lambda_{p,q}$.

The matrices $V_{p,q}$ and $\Lambda_{p,q}$ can then be used to construct the matrices $V, \Lambda \in \mathbb{C}^{3N_xN_y\times 3N_xN_y}$ respectively, that allow K to be diagonalised. We construct V and Λ in the same way as we constructed the matrix K,

$$\{V\}_{(k-1)3N_x+(j-1)3+r,(q-1)3N_x+(p-1)3+s} = \{V_{p,q}\}_{r,s}\delta_{k,q}\delta_{j,p}, \tag{6.68}$$

$$\{\Lambda\}_{(k-1)3N_x+(j-1)3+r,(q-1)3N_x+(p-1)3+s} = \{\Lambda_{p,q}\}_{r,s}\delta_{k,q}\delta_{j,p}. \tag{6.69}$$

This gives the matrices a block diagonal structure using 3×3 blocks. As the matrices $\Lambda_{p,q}$ are diagonal, Λ is a diagonal matrix. As a result, we diagonalise M such that $M = XV\Lambda V^{-1}X^*$. The matrix XV then contains the eigenvectors corresponding to the eigenvalues along the main diagonal of Λ, for the matrix M.

Now we have defined the finite difference schemes we will consider for numerically solving the 2D linearised shallow water problem, we wish to see how aliasing impacts the propagation of each wavenumber component of the solution. Once this has been achieved, we can utilise it in attempting to define numerical dissipation and dispersion for a multivariate system of PDEs.

6.2.2 Aliasing and the Poisson summation for the 2D linearised shallow water problem

The 2D linearised shallow water equations in (6.14) are posed in two dimensions. This means that the analysis in Section 5.3.1, on the effects of aliasing in two dimensions, is also relevant to this problem. In Section 6.2.1, we saw that by applying X^* to $\mathbf{Z}^0$, that we create the vector of coefficients in (6.60). This vector contains the coefficients for the 2D discrete Fourier series of the state vectors $\mathbf{U}^0$, $\mathbf{V}^0$ and $\mathbf{H}^0$. Provided the functions $u_0(x,y)$, $v_0(x,y)$ and $h_0(x,y)$ have convergent Fourier series and are continuous at every sample point, the 2D Poisson summation can be used to construct these coefficients,

$$\begin{bmatrix} \mathcal{F}_{p,q}(\mathbf{U}^0) \\ \mathcal{F}_{p,q}(\mathbf{V}^0) \\ \mathcal{F}_{p,q}(\mathbf{H}^0) \end{bmatrix} = \sqrt{N_x N_y} \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \mathbf{c}_{p-1+jN_x,q-1+kN_y}, \tag{6.70}$$

for $p = 1, \ldots, N_x$ and $q = 1, \ldots, N_y$. Here $\mathbf{c}_{j,k}$ is the coefficient for the (j,k)th wavenumber component of the Fourier series for $\mathbf{w}_0(x,y)$, defined in (6.35).

Equation (6.64) has demonstrated that it is the (p,q)th amplification matrix, that propagates this vector of coefficients, forward Δt in time. Therefore, if the functions $u_0(x,y)$, $v_0(x,y)$ and $h_0(x,y)$ have convergent Fourier series and are continuous at every sample point, we can write equation (6.64) as,

$$\begin{bmatrix} \mathcal{F}_{p,q}(\mathbf{U}^1) \\ \mathcal{F}_{p,q}(\mathbf{V}^1) \\ \mathcal{F}_{p,q}(\mathbf{\Phi}^1) \end{bmatrix} = K_{p,q}\sqrt{N_x N_y} \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \mathbf{c}_{p-1+jN_x,q-1+kN_y}. \tag{6.71}$$

This allows us to see how aliasing results in the (p,q)th amplification matrix propagating the Fourier coefficients $\mathbf{c}_{p-1+jN_x,q-1+kN_y}$, for all $j,k \in \mathbb{Z}$, for some $p = 1, \ldots, N_x$ and $q = 1, \ldots, N_y$.

In the next Section we make use of the Poisson summation to create a Fourier series representation for the numerical solution, generated by our considered finite difference schemes. This will aid us in our attempt to define definitions for numerical dissipation and dispersion in a multivariate system.

6.3 Numerical dissipation and dispersion for the 2D linearised shallow water problem

In Section 3.5, we defined Definitions 3.4 and 3.5 for numerical dissipation and numerical dispersion respectively, in problems where we wish to solve for a single variable. We now require a definition for numerical dissipation and numerical dispersion for a multivariate system. Definitions 3.4 and 3.5 were defined based on Fourier series solutions. We will again take this approach for the linearised shallow water problem.

In order to develop our definitions for numerical dissipation and dispersion for our system of equations, we need to pose both the analytical solution and the numerical solution in Fourier series form. Once we have this, we can compare how the two Fourier series progress their solution forward Δt in time. This is same method as we used in Chapters 3 and 5, to define the functions $g(\cdot)$ and $g^{scheme}(\cdot)$, that we used to define Definitions 3.4 and 3.5.

In Section 6.1, we developed the analytical solution to the 2D linearised shallow water problem, as a Fourier series solution. This is found in Equation (6.43) and provides a vector solution constructed of the three variables we are interested in. Examining this solution we see that instead of an eigenvalue determining the magnitude and phase change to each wavenumber component of the solution, we have a 3×3 matrix. The matrix applying the phase and amplitude changes to the (p, q)th wavenumber component in time t, is $\hat{A}e^{\hat{D}_{p,q}t}\hat{A}^{-1}$, for $p, q \in \mathbb{Z}$. Define the function $\mathbf{b}_{p,q} : [0, \infty) \to \mathbb{C}^3$, such that $t \mapsto \mathbf{b}_{p,q}(t) = \hat{A}e^{\hat{D}_{p,q}t}\hat{A}^{-1}\mathbf{c}_{p,q}$, which defines the Fourier coefficient for the analytical solution at time t. As with the 2D linear advection problem, we can define the function $\mathbf{g}_{p,q} : \mathbb{C}^3 \to \mathbb{C}^3$ such that,

$$\mathbf{z} \mapsto \mathbf{g}_{p,q}(\mathbf{z}) = \hat{A}e^{\hat{D}_{p,q}\Delta t}\hat{A}^{-1}\mathbf{z}. \tag{6.72}$$

This maps the vector of coefficients of the (p, q)th wavenumber component, Δt through time, $\mathbf{g}_{p,q}(\mathbf{b}_{p,q}(t)) = \mathbf{b}_{p,q}(t + \Delta t)$.

We now construct a Fourier series representation for the outcome of our considered finite difference schemes. Consider the Fourier series for the initial condition in (6.34). Equation (6.71) shows that the Upwind and Crank-Nicolson finite difference schemes, create a numerical solution to the linearised shallow water equations every Δt in time, by post-multiplying Fourier coefficient $\mathbf{c}_{p-1+jN_x,q-1+kN_y}$ by $K_{p,q}$ for all $p = 1, \ldots, N_x$, $q = 1, \ldots, N_y$ and $j, k \in \mathbb{Z}$. Therefore a Fourier series representation of the numerical solution is given by the function $\mathbf{a} : \mathbb{R} \times \mathbb{R} \times [0, \infty) \to \mathbb{R}^3$ such that,

$$(x, y, t) \mapsto \mathbf{a}(x, y, t) = \sum_{p=-\infty}^{\infty} \sum_{q=-\infty}^{\infty} K_{[p]_{N_x}+1,[q]_{N_y}+1}^{\frac{t}{\Delta t}} \mathbf{c}_{p,q} e^{2\pi ipx} e^{2\pi iqy}, \tag{6.73}$$

provided $\mathbf{w}_0(x, y)$ has a convergent Fourier series. In this instance the (p, q)th wavenumber component is propagated by the amplification matrix $K_{[p]_{N_x}+1,[q]_{N_y}+1}$ for $p, q \in \mathbb{Z}$.

This Fourier series forms an approximation for the analytical solution in equation (6.43). Evaluating (6.73) at non-integer multiples of Δx and Δy, interpolates the numerical solution in space. Similarly, evaluating at times that are non-integer multiples of Δt, interpolates the numerical solution in time. Then we can define the function $\mathbf{g}_{p,q}^{scheme} : \mathbb{C}^3 \to \mathbb{C}^3$ that maps the coefficient for the (p, q) wavenumber component Δt through time,

$$\mathbf{z} \mapsto \mathbf{g}_{p,q}^{scheme}(\mathbf{z}) = K_{[p]_{N_x}+1,[q]_{N_y}+1}\mathbf{z} \tag{6.74}$$

Now we have the functions $g_{p,q}(\mathbf{z})$ and $g^{scheme}_{p,q}(\mathbf{z})$, we can identify the amplification matrix for the (p,q)th wavenumber component, in the analytical and numerical solution respectively. It is these matrices we wish to compare to identify the numerically dissipative and dispersive properties of the considered finite difference schemes. However, it is not possible to obtain these matrices by evaluating the functions at a given value of $\mathbf{z}$. Therefore we take the amplifications matrices directly from the functions.

The amplification matrix for the (p,q)th wavenumber component, obtained from the function $\mathbf{g}_{p,q}(\mathbf{z})$ is,

$$\hat{A}e^{\hat{D}_{p,q}\Delta t}\hat{A}^{-1} = \hat{A}E_{p,q}e^{\Gamma_{p,q}\Delta t}E^{*}_{p,q}\hat{A}^{-1}. \tag{6.75}$$

We see that this matrix has been diagonalised, using the eigenvectors found in the columns of the matrix $\hat{A}E_{p,q}$. The eigenvalues of the matrix are found in $e^{\Gamma_{p,q}\Delta t}$.

The amplification matrix for the (p,q)th wavenumber component, obtained from the function $\mathbf{g}^{scheme}_{p,q}(\mathbf{z})$, is $K_{[p]_{N_x}+1,[q]_{N_y}+1}$. If we change the amplification matrix $K_{[p]_{N_x}+1,[q]_{N_y}+1}$ into the same basis that diagonalises (6.75), we could potentially compare the eigenvalues of the amplification matrices, to determine the numerically dissipative and dispersive properties of the scheme. However, this requires that the basis that diagonalises $\hat{A}e^{\hat{D}_{p,q}\Delta t}\hat{A}^{-1}$ also diagonalise $K_{[p]_{N_x}+1,[q]_{N_y}+1}$ for each $p,q \in \mathbb{Z}$. When this occurs, the matrices are termed simultaneously diagonalisable.

Two matrices are simultaneously diagonalisable if and only if the two matrices are commutable [93]. Therefore, we require that,

$$\hat{A}e^{\hat{D}_{p,q}\Delta t}\hat{A}^{-1}K_{[p]_{N_x}+1,[q]_{N_y}+1} = K_{[p]_{N_x}+1,[q]_{N_y}+1}\hat{A}e^{\hat{D}_{p,q}\Delta t}\hat{A}^{-1}.$$

If we calculate this for the Upwind and Crank-Nicolson schemes, we find that $K_{[p]_{N_x}+1,[q]_{N_y}+1}$ and $\hat{A}e^{\hat{D}_{p,q}\Delta t}\hat{A}^{-1}$ do not commute for either scheme, for all $p,q \in \mathbb{Z}$. As a result, it is not possible to diagonalise the (p,q)th amplification matrix of the Upwind and Crank-Nicolson schemes, using the eigenbasis of $\hat{A}e^{\hat{D}_{p,q}\Delta t}\hat{A}^{-1}$ for each $p,q \in \mathbb{Z}$.

Therefore, we cannot compare the eigenvalues of these two matrices as we had desired. We are left with requiring a method for comparing the amplification matrix $K_{[p]_{N_x}+1,[q]_{N_y}+1}$ for the scheme, with the matrix $\hat{A}e^{\hat{D}_{p,q}\Delta t}\hat{A}^{-1}$, for each $p,q \in \mathbb{Z}$. The question that now arises is, *how do you extend the definitions for numerical dissipation and dispersion to multivariate systems?* Examining the literature, we have not been able to identify a definition for numerical dissipation and dispersion for multivariate systems, whose amplification matrices are not simultaneously diagonalisable. Our aim is to answer this question in the following Sections.

6.3.1 A strict interpretation

The amplification matrices of the finite difference schemes directly propagate the resolvable wavenumber coefficients of the numerical solution. Therefore we will initially

try to define numerical dissipation and dispersion for the resolvable wavenumber components. Hence we initially consider the amplification matrices $\hat{A}e^{\hat{D}_{p-1,q-1}\Delta t}\hat{A}^{-1}$ and $K_{p,q}$, for $p = 1, \ldots, \frac{N_x+1}{2}$ and $q = 1, \ldots, \frac{N_y+1}{2}$. In order to compare the effects of these matrices on the vector of coefficients $[\mathcal{F}_{p,q}(\mathbf{U}^n), \mathcal{F}_{p,q}(\mathbf{V}^n), \mathcal{F}_{p,q}(\mathbf{H}^n)]^T$, we will compare them in the same basis. We will choose the basis which diagonalises $\hat{A}e^{\hat{D}_{p-1,q-1}\Delta t}\hat{A}^{-1}$.

The matrix $\hat{A}e^{\hat{D}_{p-1,q-1}\Delta t}\hat{A}^{-1}$ diagonalises in the basis given by the columns of the matrix, $\hat{A}E_{p-1,q-1}$. Therefore making a change to this basis results in $\hat{A}e^{\hat{D}_{p-1,q-1}\Delta t}\hat{A}^{-1}$ becoming a diagonal amplification matrix, $e^{\Gamma_{p-1,q-1}\Delta t}$. In this new basis, the amplification matrix $K_{p,q}$ becomes, $E^*_{p-1,q-1}\hat{A}^{-1}K_{p,q}\hat{A}E_{p-1,q-1}$. This is a full matrix, as it is not simultaneously diagonalisable with $\hat{A}e^{\hat{D}_{p-1,q-1}\Delta t}\hat{A}^{-1}$.

Consider applying the matrix $e^{\Gamma_{p-1,q-1}\Delta t}$, to our vector of coefficients, in our new basis. Let the vector $[\hat{\alpha}_{p-1,q-1}, \hat{\beta}_{p-1,q-1}, \hat{\gamma}_{p-1,q-1}]^T$ denote the vector of coefficients in the new basis, $\hat{\alpha}_{p-1,q-1}, \hat{\beta}_{p-1,q-1}, \hat{\gamma}_{p-1,q-1} \in \mathbb{C}$ and $(\lambda_m)_{p-1,q-1} \in \mathbb{C}$ denote the mth eigenvalue of $\hat{A}e^{\hat{D}_{p-1,q-1}\Delta t}\hat{A}^{-1}$, $m = 1, 2, 3$. Then,

$$e^{\Gamma_{p-1,q-1}\Delta t}\begin{bmatrix}\hat{\alpha}_{p,q}\\ \hat{\beta}_{p,q}\\ \hat{\gamma}_{p,q}\end{bmatrix} = \begin{bmatrix}(\lambda_1)_{p-1,q-1}\hat{\alpha}_{p,q}\\ (\lambda_2)_{p-1,q-1}\hat{\beta}_{p,q}\\ (\lambda_3)_{p-1,q-1}\hat{\gamma}_{p,q}\end{bmatrix} \tag{6.76}$$

Then the left-hand side of this equation is what we wish to attain through

$$E^*_{p-1,q-1}\hat{A}^{-1}K_{p,q}\hat{A}E_{p-1,q-1}\begin{bmatrix}\hat{\alpha}_{p,q}\\ \hat{\beta}_{p,q}\\ \hat{\gamma}_{p,q}\end{bmatrix}. \tag{6.77}$$

Suppose we interpret Definitions 3.4 and 3.5 strictly. Then we could term the scheme with amplification matrix $K_{p,q}$, numerically dissipative when (6.77) results in an entry in the resultant vector, which has an incorrect amplitude. Similarly, we could term the scheme numerically dispersive when (6.77) results in an entry in the resultant vector, which has an incorrect phase.

Suppose we trial this interpretation of the definitions for numerical dissipation and dispersion. Consider,

$$\begin{bmatrix}a_1 & b_1 & c_1\\ \star & b_2 & \star\\ \star & \star & c_3\end{bmatrix}\begin{bmatrix}\hat{\alpha}_{p,q}\\ \hat{\beta}_{p,q}\\ \hat{\gamma}_{p,q}\end{bmatrix} = \begin{bmatrix}s_1\\ s_2\\ s_3\end{bmatrix}. \tag{6.78}$$

Here $\star$ denotes a potentially non-zero entry in the matrix, which we do not require for our calculations. Let the matrix on the left-hand side represent the matrix

$$E^*_{p-1,q-1}\hat{A}^{-1}K_{p,q}\hat{A}E_{p-1,q-1}.$$

The vector on the right-hand side, is the result of the matrix-vector multiplication. The

scheme will be considered numerically non-dissipative if $|s_m| = |(\lambda_m)_{p-1,q-1}||\hat{\alpha}_{p,q}|$ and numerically non-dispersive if $\text{phase}(s_m) = \text{phase}((\lambda_m)_{p-1,q-1}\hat{\alpha}_{p,q})$ for all $m = 1, 2, 3$.

These definitions for numerical dissipation and dispersion, must hold for any choice of $\hat{\alpha}_{p,q}$, $\hat{\beta}_{p,q}$ and $\hat{\gamma}_{p,q}$. The first line of (6.78) gives that,

$$\begin{aligned} & a_1\hat{\alpha}_{p,q} + b_1\hat{\beta}_{p,q} + c_1\hat{\gamma}_{p,q} = (\lambda_1)_{p-1,q-1}\hat{\alpha}_{p,q}, \\ \Rightarrow \quad & [a_1 - (\lambda_1)_{p-1,q-1}]\hat{\alpha}_{p,q} + b_1\hat{\beta}_{p,q} + c_1\hat{\gamma}_{p,q} = 0. \end{aligned}$$

As this must hold for any $\hat{\alpha}_{p,q}$, $\hat{\beta}_{p,q}$ and $\hat{\gamma}_{p,q}$, we have that $a_1 = (\lambda_1)_{p-1,q-1}$ and $b_1 = c_1 = 0$. Similarly, when considering (6.78) for s_2 and s_3, we find that for $K_{p,q}$ to be either numerically non-dissipative and/or non-dispersive, we require that the matrix in (6.78) be diagonal.

Therefore the scheme is numerically non-dissipative when $K_{p,q}$ is a diagonal matrix with $|(a_1)_{p,q}| = |(\lambda_1)_{p-1,q-1}|$, $|(b_2)_{p,q}| = |(\lambda_2)_{p-1,q-1}|$ and $|(c_3)_{p,q}| = |(\lambda_3)_{p-1,q-1}|$. Similarly the scheme is non-dispersive when $K_{p,q}$ is a diagonal matrix with $\text{phase}((a_1)_{p,q}) = \text{phase}((\lambda_1)_{p,q})$, $\text{phase}((b_2)_{p,q}) = \text{phase}((\lambda_2)_{p,q})$ and $\text{phase}((c_3)_{p,q}) = \text{phase}((\lambda_3)_{p,q})$.

This seems a rather restrictive definition as it requires that the amplification matrix $K_{p,q}$ of the considered finite difference scheme, be simultaneously diagonalisable with $\hat{A}e^{\hat{D}_{p-1,q-1}\Delta t}\hat{A}^{-1}$. As a result, our considered schemes cannot be classed as either numerically non-dissipative or non-dispersive with respect to the resolvable wavenumber components for $p = 1, \ldots, \frac{N_x+1}{2}$ and $q = 1, \ldots, \frac{N_y+1}{2}$. A similar outcome will arise when performing the same analysis for the other resolvable wavenumber components. Therefore, in the following Section, we trial an alternative definition. We define numerical dissipation and dispersion through the *Polar Decomposition* of the amplification matrices, since this extends the polar co-ordinate form from scalars to matrices.

6.3.2 The polar decomposition

The polar decomposition of a matrix, is the matrix form of writing scalars in polar co-ordinate form. Consider the amplification matrix $K_{p,q}$ for $p = 1, \ldots, N_x$ and $q = 1, \ldots, N_y$. Then the polar decomposition of this matrix is given by, $K_{p,q} = P_{p,q}U_{p,q}$ [95] where,

- $U_{p,q} \in \mathbb{C}^{3\times 3}$ is a rotation matrix [96],
- $P_{p,q} \in \mathbb{R}^{3\times 3}$ is a Hermitian positive definite matrix [95], applying a stretch [96].

If the matrix $K_{p,q}$ is invertible, then the polar decomposition is unique. The matrix $P_{p,q}$ is calculated by $P_{p,q} = \sqrt{K^*_{p,q}K_{p,q}}$. Once this has been calculated, we can calculate $U_{p,q} = P^{-1}_{p,q}K_{p,q}$ [95].

If we take the polar decomposition of $\hat{A}e^{\hat{D}_{p-1,q-1}\Delta t}\hat{A}^{-1}$ for $p, q \in \mathbb{Z}$, we find that the matrix is already in polar decomposition form. This is due to the fact that $\hat{A}e^{\hat{D}_{p-1,q-1}\Delta t}\hat{A}^{-1}$ is a rotation matrix, as discussed in Section 6.1.2.

Suppose instead of considering $K_{p,q}$ directly, we consider the matrix,

$$B_{p,q} := \tilde{K}_{p.q}^{-1} K_{p,q} = \hat{A} e^{-\hat{D}_{p-1,q-1}\Delta t} \hat{A}^{-1} K_{p,q}, \tag{6.79}$$

where $\tilde{K}_{p,q}$ is the amplification matrix of the analytical solution which propagates the (p,q)th resolvable wavenumber component for $p = 1, \ldots, N_x$ and $q = 1, \ldots, N_y$. These are the amplification matrices of the MNIMC scheme for the 2D linearised shallow water problem, defined in the next section. When $p = 1, \ldots, \frac{N_x+1}{2}$ and $q = 1, \ldots, \frac{N_y+1}{2}$,

$$\tilde{K}_{p,q} = \hat{A} e^{\hat{D}_{p-1,q-1}\Delta t} \hat{A}^{-1}.$$

The matrix $B_{p,q}$ is the (p,q)th amplification matrix of the matrix $\tilde{M}^{-1}M$. Here $\tilde{M}$ and M are the matrices implementing the MNIMC and the considered imperfect scheme, respectively. The matrix $\tilde{M}^{-1}M$ allows the imperfect scheme to move the state of the system forward Δt in time and the MNIMC scheme then moves the system Δt backwards through time. We choose to invert the matrix implementing the MNIMC scheme as we know this is always invertible. Through examining $\tilde{M}^{-1}M$, we aim to examine the ability of the matrix M to correctly propagate the resolvable wavenumber components of the numerical solution. If M introduces no numerical model error into the resolvable wavenumber components, then we would expect $\tilde{M}^{-1}M = I_{3N_xN_y}$, the $3N_xN_y \times 3N_xN_y$ identity matrix ie: $M = \tilde{M}$.

Consider the matrix $B_{p,q}$ in the basis that diagonalises $\tilde{K}_{p,q}$. This is the matrix,

$$E_{p-1,q-1}^{*} \hat{A}^{-1} B_{p,q} \hat{A} E_{p-1,q-1} = e^{-\Gamma_{p-1,q-1}\Delta t} E_{p-1,q-1}^{*} \hat{A} K_{p,q} \hat{A}^{-1} E_{p-1,q-1}, \tag{6.80}$$

when $p = 1, \ldots, \frac{N_x+1}{2}$ and $q = 1, \ldots, \frac{N_y+1}{2}$. Performing the polar decomposition of this matrix may yield a way to define the equivalent of numerical dissipation and dispersion for multivariate systems of PDEs. Developing a way of testing this is left as future work.

The idea behind this can be seen by considering the 1D linear advection problem, where $N = N_x$. Consider the matrix $\tilde{M}^{-1}M$, where $\tilde{M} \in \mathbb{R}^{N_x \times N_x}$ and $M \in \mathbb{R}^{N_x \times N_x}$ are the matrices implementing the MNIMC and the considered imperfect scheme, for the 1D linear advection problem in (3.1). This uses the same notation and definitions as defined in Chapter 3. The 1D DFT basis diagonalises both $\tilde{M}$ and M, so we consider $\tilde{M}^{-1}M$ in this basis. This matrix is $\tilde{\Lambda}^{-1}\Lambda$, where $\tilde{\Lambda}$ and Λ are the diagonal matrices, with the eigenvalues of the MNIMC and the considered imperfect scheme along their main diagonals respectively. Let the pth eigenvalue of $\tilde{\Lambda}$ and Λ, be defined by $\tilde{\lambda}_p = |\tilde{\lambda}_p| e^{i\tilde{\theta}_p}$ and $\lambda_p = |\lambda_p| e^{i\theta_p}$ respectively, for $\tilde{\theta}_p, \theta_p \in [-\pi, \pi)$ and $p = 1, \ldots, N_x$. Then the pth eigenvalue of $\tilde{\Lambda}^{-1}\Lambda$ is,

$$\frac{\lambda_p}{\tilde{\lambda}_p} = |\lambda_p| e^{-i\phi_p}, \tag{6.81}$$

where $\phi_p = \tilde{\theta}_p - \theta_p$ for $p = 1, \ldots, N$. Examining the magnitude and phase of (6.81), reveals the numerically dissipative and numerically dispersive properties of λ_p. If $|\lambda_p| = 1$, then λ_p is numerically non-dissipative with respect to the pth resolvable wavenumber component. If $\phi_p = 0$, then the scheme is numerically non-dispersive with respect to the pth resolvable wavenumber component.

We now return our attention to the 2D linearised shallow water problem. The formulation presented in this Section, for forming an equivalent definition for numerical dissipation and dispersion in systems of PDEs, needs to be tested to check its suitability. In the next Section, we focus on the problems associated with generating perfect observations of the 2D linearised shallow water problem, for use in our strong constraint 4D-Var data assimilation problem. This includes the construction of the MNIMC scheme for this problem.

6.4 Generating perfect observations

Generating perfect observations is a challenge for the 2D linearised shallow water problem in (6.14)-(6.16). In the case of the 1D and 2D linear advection problems, we were able to make use of the *circshift* function in MATLAB ®[74], to numerically generate perfect observations. In the case of this problem, there is no easy way to generate perfect observations numerically or algebraically. Perfect observations cannot be generated numerically from the Fourier series solution to the problem. The only possible way is to create the MNIMC scheme for this problem and see if it has a shifted periodic nature, as demonstrated for the MNIMC schemes for the 1D and 2D linear advection problems. It can then be used to generate perfect observations as described in Section 3.7.2. However this will be computationally expensive and will limit the dimension of the problem we can consider.

6.4.1 The MNIMC scheme for the 2D linearised shallow water problem

The MNIMC scheme for the 2D linearised shallow water problem, can be developed using the theory developed in Section 5.6.1 for the 2D linear advection problem, as both systems are defined in 2D, using Fourier series. The difference between the two systems is that the 2D linear advection problem solves for one variable whilst the 2D linearised shallow water problem solves for three variables. The result is that for the 2D linear advection problem, eigenvalues for the scheme need to be chosen whilst for the 2D linearised shallow water problem, amplification matrices need to be chosen.

As for the 1D and 2D linear advection problems, we construct the MNIMC scheme using the amplification matrices which correspond to the resolvable wavenumber components of the analytical solution. As problem (6.14)-(6.16) is defined in 2D, we can choose the amplification matrix for each resolvable wavenumber component us-

ing (5.39), but use the amplification matrix from $\mathbf{g}_{p,q}(\mathbf{z})$ in equation (6.72).

Denote the (p,q)th amplification matrix for the MNIMC scheme, by $\tilde{K}_{p,q} \in \mathbb{C}^{3\times 3}$, for $p = 1,\ldots,N_x$ and $q = 1,\ldots,N_y$. As a finite difference scheme creates a discrete Fourier series solution, we require that given any amplification matrix of the scheme, its complex conjugate is an amplification matrix for the complex conjugate wavenumber component of the scheme. This can only occur when both N_x and N_y are odd, as demonstrated by Section 5.6.1. Therefore we will only consider the MNIMC scheme for this problem in this situation. Then we choose the amplification matrices for the scheme using (5.39),

$$\tilde{K}_{p,q} = \begin{cases} \hat{A}e^{\hat{D}_{p-1,q-1}\Delta t}\hat{A}^{-1}, & \text{for } p = 1,\ldots,\frac{N_x+1}{2},\ q = 1,\ldots,\frac{N_y+1}{2}, \\ \hat{A}e^{\hat{D}_{p-1,-N_y+q-1}\Delta t}\hat{A}^{-1}, & \text{for } p = 1,\ldots,\frac{N_x+1}{2}, \text{ and } q = \frac{N_y+3}{2},\ldots,N_y, \\ \hat{A}e^{\hat{D}_{-N_x+p+1,q-1}\Delta t}\hat{A}^{-1}, & \text{for } p = \frac{N_x+3}{2},\ldots,N_x,\ q = 1,\ldots,\frac{N_y+1}{2}, \\ \hat{A}e^{\hat{D}_{-N_x+p-1,-N_y+q+1}\Delta t}\hat{A}^{-1}, & \text{for } p = \frac{N_x+3}{2},\ldots,N_x, \text{ and } q = \frac{N_y+3}{2},\ldots,N_y. \end{cases} \tag{6.82}$$

We now define the block diagonal matrix $\tilde{K} \in \mathbb{C}^{3N_xN_y\times 3N_xN_y}$, such that the amplification matrices $\tilde{K}_{p,q}$, form the blocks along its main diagonal, similarly to K in (6.61) such that,

$$\{\tilde{K}\}_{(k-1)3N_x+(j-1)3+r,(q-1)3N_x+(p-1)3+s} = \{\tilde{K}_{p,q}\}_{r,s}\delta_{k,q}\delta_{j,p}. \tag{6.83}$$

Then the matrix $\tilde{M} \in \mathbb{R}^{3N_xN_y\times 3N_xN_y}$ implementing the MNIMC scheme for the 2D linearised shallow water problem is defined by $\tilde{M} = X\tilde{K}X^*$. Define the numerical solution generated by the MNIMC scheme at time $n\Delta t$ by $\tilde{\mathbf{Z}}^n \in \mathbb{R}^{3N_xN_y}$ such that $\tilde{\mathbf{Z}}^{n+1} = \tilde{M}\tilde{\mathbf{Z}}^n$, for $n \in \mathbb{N}_0$, where $\tilde{\mathbf{Z}}^0$ is constructed by sampling the initial conditions $w(x,y,0)$. This vector then contains the numerical solutions to the 2D linearised shallow water problem, generated by the MNIMC scheme, with the same structure as $\mathbf{Z}^n$ for the Upwind and Crank-Nicolson schemes. Denote the numerical solutions stacked in this matrix by $\tilde{\mathbf{W}}^n_{j,k} \in \mathbb{C}^3$ such that,

$$\tilde{\mathbf{W}}^n_{j,k} = \begin{bmatrix} \tilde{U}^n_{j,k} \\ \tilde{V}^n_{j,k} \\ \tilde{H}^n_{j,k} \end{bmatrix}. \tag{6.84}$$

The eigenvalues of $\tilde{M}$, are given by the eigenvalues of $\tilde{K}_{p,q}$, for $p = 1,\ldots,N_x$ and $q = 1,\ldots,N_y$. Since the eigenvalues of these matrices always have unit magnitude, the scheme is always numerically stable. The consistency of the scheme is proved in the following Lemma.

Lemma 6.1. *Suppose the initial conditions $u_0(x,y)$, $v_0(x,y)$ and $h_0(x,y)$ for problem*

(6.14)-(6.16) are multiplicatively separable functions such that $u_0(x,y) = \hat{u}_1(x)\hat{u}_2(y)$, $v_0(x,y) = \hat{v}_1(x)\hat{v}_2(y)$ *and* $h_0(x,y) = \hat{h}_1(x)\hat{h}_2(y)$, *where* $\hat{u}_1, \hat{u}_2, \hat{v}_1, \hat{v}_2, \hat{h}_1, \hat{h}_2 : \mathbb{R} \to \mathbb{R}$ *such that* $x \mapsto \hat{u}_1(x), \hat{v}_1(x), \hat{h}_1(x)$ *and* $y \mapsto \hat{u}_2(y), \hat{v}_2(y), \hat{h}_2(y)$ *and all have convergence Fourier series. Let* $r_{1u}, r_{2u}, r_{1v}, r_{2v}, r_{1h}, r_{2h} \in \mathbb{N}_0$ *denote the regularities of* $\hat{u}_1(x)$, $\hat{u}_2(y)$, $\hat{v}_1(x)$, $\hat{v}_2(y)$, $\hat{h}_1(x)$ *and* $\hat{h}_2(y)$ *over* $(0,1)$ *respectively and define* $r_a := \min\{r_{1u}, r_{1v}, r_{1h}\}$ *and* $r_b := \min\{r_{2u}, r_{2v}, r_{2h}\}$.

Also let the assumptions of Section 6.2, that allow the MNIMC scheme to be defined as in Section 6.4.1, hold true. Set the CFL number $h \in \mathbb{R}^+$ *to be a fixed constant. Then the truncation error for the MNIMC scheme is such that,*

$$\left\|\boldsymbol{\tau}_{j,k}^{n+1}\right\|_2 = \mathcal{O}\left(\Delta x^{r_a}\Delta y^{r_b}\right) + \mathcal{O}\left(\Delta x^{r_a}\right) + \mathcal{O}\left(\Delta y^{r_b}\right), \tag{6.85}$$

for all $j = 0, \ldots, N_x - 1$, $k = 0, \ldots, N_y - 1$ *and* $n \in \mathbb{N}_0$. *Then for sufficiently smooth functions such that* $r_a, r_b \in \mathbb{N}$,

$$\tau_{j,k}^{n+1} \to 0 \text{ and } \Delta t \to 0 \text{ as } \Delta x, \Delta y \to 0,$$

for all $j = 0, \ldots, N_x - 1$, $k = 0, \ldots, N_y - 1$ *and* $n \in \mathbb{N}_0$.

Proof. The analytical solution to the 2D linearised shallow water problem in (6.14)-(6.16), is a vector in $\mathbb{R}^3$, so the truncation error is a vector in $\mathbb{R}^3$. Define $\boldsymbol{\tau}_{j,k}^{n+1} \in \mathbb{R}^3$ to be the error between $\tilde{\mathbf{W}}_{j,k}^{n+1}$ and the state of the system when $\tilde{M}$ is used to propagate $\tilde{\mathbf{W}}_{j,k}^{n}$ forward Δt in time, for all $j = 0, \ldots, N_x - 1$, $k = 0, \ldots, N_y - 1$ and $n \in \mathbb{N}_0$.

Section 5.4 defines Fourier series representations for the analytical solution in $\tilde{\mathbf{W}}_{j,k}^{n+1}$ and the numerical solution $\mathbf{W}_{j,k}^{n}$. These expansions are equal to the quantity they represent as the initial conditions $u_0(x,y)$, $v_0(x,y)$ and $h_0(x,y)$ have convergent Fourier series and they are multiplicatively separable functions, constructed from continuous functions. Then using these Fourier series, we obtain,

$$\boldsymbol{\tau}_{j,k}^{n+1} = \sum_{p=-\infty}^{\infty}\sum_{q=-\infty}^{\infty} \hat{A}\left(e^{\hat{D}_{p,q}\Delta t} - e^{\hat{D}_{[p]_{N_x}+1,[q]_{N_y}+1}\Delta t}\right) e^{\hat{D}_{p,q}n\Delta t}\hat{A}^{-1} e^{\frac{2\pi i p j}{N_x}} e^{\frac{2\pi i q k}{N_y}}. \tag{6.86}$$

Taking the l_2-norm, applying the triangle inequality and using the fact that $\|\mathbf{z}\|_2 \leq \|\mathbf{z}\|_1$ for all $\mathbf{z} \in \mathbb{R}^3$,

$$\left\|\boldsymbol{\tau}_{j,k}^{n+1}\right\|_2 = \max\left\{\frac{\sqrt{\phi}}{g}, \frac{g}{\sqrt{\phi}}\right\} \sum_{p=-\infty}^{\infty}\sum_{q=-\infty}^{\infty} \left\|e^{\hat{D}_{p,q}\Delta t} - e^{\hat{D}_{[p]_{N_x}+1,[q]_{N_y}+1}\Delta t}\right\|_2 \left\|\mathbf{c}_{p,q}\right\|_1, \tag{6.87}$$

as $\left\|e^{\hat{D}_{j,k}\Delta t}\right\|_2 = 1$ and $\left\|\hat{A}\right\|_2\left\|\hat{A}^{-1}\right\|_2 = \max\left\{\frac{g}{\sqrt{\phi}}, \frac{\sqrt{\phi}}{g}\right\}$ by direct calculation.

Define $\mathbf{c}_{p,q} = [u_{p,q}, v_{p,q}, h_{p,q}]^T$, where $u_{p,q}, v_{p,q}, h_{p,q}$ are the Fourier coefficients for the Fourier series of $u_0(x,y)$, $v_0(x,y)$ and $h_0(x,y)$ respectively for all $p, q \in \mathbb{Z}$. Then

by Lemma 5.8,

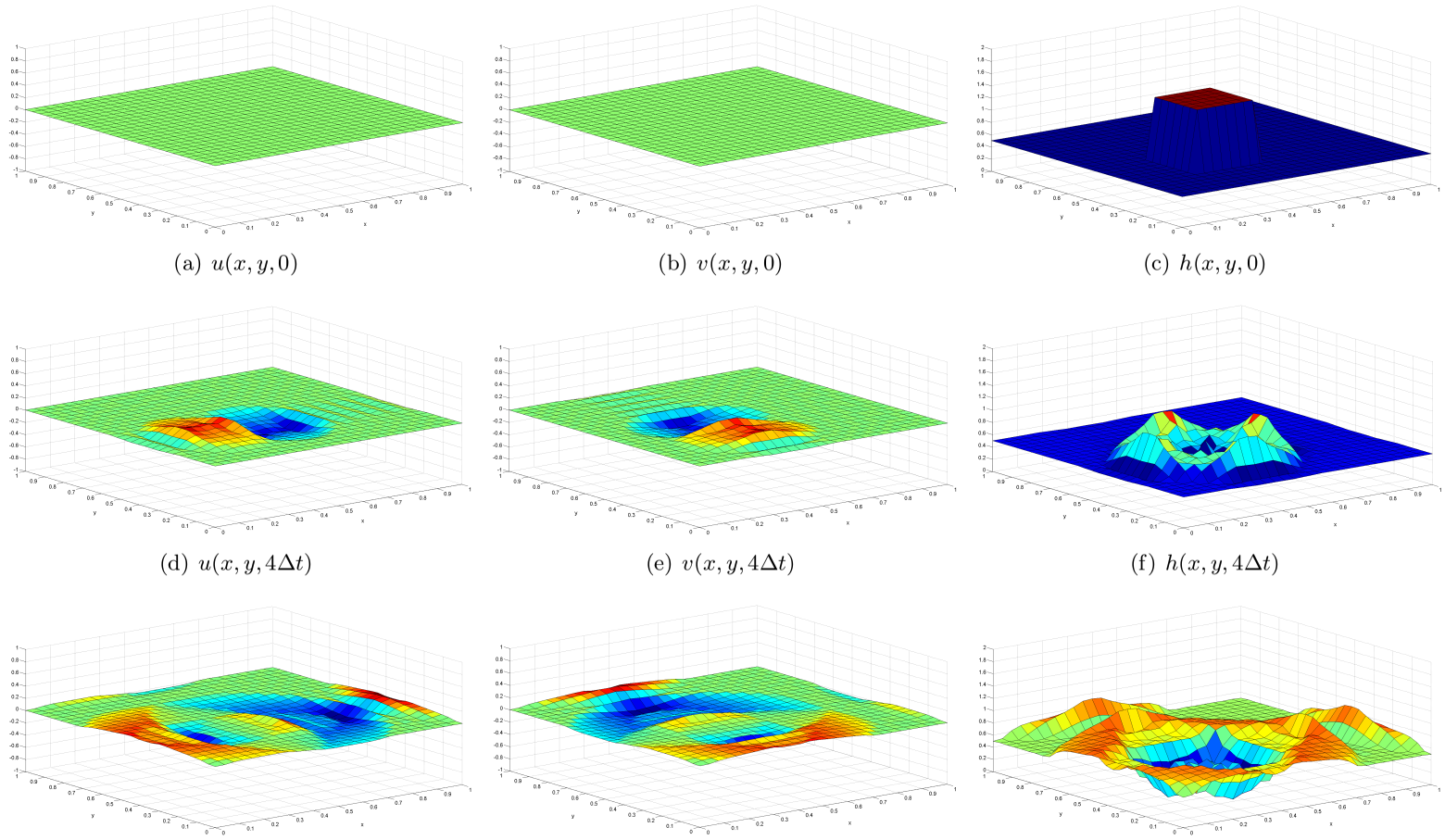
$$\begin{aligned} & \|\mathbf{c}_{p,q}\|_1 \\ = \; & |u_{p,q}| + |v_{p,q}| + |h_{p,q}|\,, \\ \leq \; & \begin{cases} U_1 + V_1 + H_1, & \text{for } p = q = 0, \\ \frac{U_2+V_2+H_2}{|q|^{r_b+1}}, & \text{for } p = 0 \text{ and } q \in \mathbb{Z} \setminus \{0\}, \\ \frac{U_3+V_3+H_3}{|p|^{r_a+1}}, & \text{for } p \in \mathbb{Z} \setminus \{0\} \text{and } q = 0, \\ \frac{U_4+V_4+H_4}{|p|^{r_a+1}|q|^{r_b+1}}, & \text{for } p, q \in \mathbb{Z} \setminus \{0\}, \end{cases} \end{aligned} \tag{6.88}$$

Here $U_d \in \mathbb{R}$ is the value of A_d in Lemma 5.8, for the function $u_0(x, y)$ for $d = 1, \ldots, 4$. Similar is true for $V_d, H_d \in \mathbb{R}$ for the functions $v_0(x, y)$ and $h_0(x, y)$ respectively, for $d = 1, \ldots, 4$. Then by following identical steps to Lemma 5.3, we obtain the result in (6.85). □

Now we have investigated the convergence properties of the MNIMC scheme for the linearised shallow water equations, we can investigate whether the aliasing error in the scheme has a shifted periodic nature. This property would allow us to generate perfect observations numerically, as discussed in Section 3.7.2. Figure 6.1 demonstrates the results of implementing the MNIMC scheme for the shallow water problem. It is not obvious from this Figure if the aliasing error in the scheme has a shifted periodic nature.

As the Rossby wave solution of the MNIMC scheme does not change with time, it solves for the Rossby wave solution exactly, if the initial conditions of the problem have convergent Fourier series. Therefore the aliasing error introduced by the MNIMC scheme into the Rossby wave solution of the scheme, is always zero. This effectively gives the aliasing error introduced into this solution, a shifted one-periodic nature, as it repeats with each application of the MNIMC scheme.

The next step is to determine if a shifted periodic nature exists in the aliasing error introduced by the MNIMC scheme into its inertia-gravity wave solutions. Lemmas 3.12 and 5.6 determine the shifted periodic nature of the MNIMC scheme for the 1D and 2D linear advection problems respectively. Whilst conducting this analysis, we found that the CFL number was of great help. Therefore in the following Section, we define the CFL number for the 2D linearised shallow water problem, before attempting to analyse the aliasing errors introduced by the MNIMC scheme for the problem.



(a) $u(x, y, 0)$ (b) $v(x, y, 0)$ (c) $h(x, y, 0)$

(d) $u(x, y, 4\Delta t)$ (e) $v(x, y, 4\Delta t)$ (f) $h(x, y, 4\Delta t)$

(g) $u(x, y, 10\Delta t)$ (h) $v(x, y, 10\Delta t)$ (i) $h(x, y, 10\Delta t)$

Figure 6.1: *The results of applying the MNIMC scheme for the linearised shallow water problem in (6.14)-(6.16), to the initial conditions* $u_0(x, y) = 0$, $v_0(x, y) = 0$ *and* $h_0(x, y)$ *defined by (5.134). The results were generated using* $N_x = N_y = 3^3$, $\Delta t = 0.01s$, $g = 9.81ms^{-1}$ *and* $H = 1$. *We also choose* $f = 10^{-4}s^{-1}$, *the value chosen by Daley [1], in his numerical experiments.*

6.4.2 The CFL number for the 2D linearised shallow water problem

In this Section, we derive the CFL number for the 2D linearised shallow water problem in (6.14)-(6.16). The following derivation of the CFL number for the considered 2D linearised shallow water problem is not known to be present in any literature, but the method used to derive it is not new. It is hoped that this quantity will aid us in determining if the aliasing errors introduced by the MNIMC scheme for this problem, have a shifted periodic nature. The form of the CFL number for a finite difference scheme, used to solve a 2D problem, was derived in Section 5.5.4. In order to define the CFL number for the 2D linearised shallow water problem, we require the propagation speed along the characteristic equations of the analytical solution. This will allow us to define the domain of dependence of the PDE and hence define the domain of dependence for the scheme, such that the former lies within the latter.

In order to define the propagation speeds along the characteristic equations for the 2D linearised shallow water problem, we need to identify the characteristic equations for the problem. As we have three solutions; the Rossby wave and the two gravity waves, there will be a different characteristic equation for each solution. Each solution is constant along its corresponding characteristic equation [14]. Therefore we require that,

$$\frac{d\mathbf{w}}{ds} = \mathbf{0} \Rightarrow \frac{\partial \mathbf{w}}{\partial x}\frac{dx}{ds} + \frac{\partial \mathbf{w}}{\partial y}\frac{dy}{ds} + \frac{\partial \mathbf{w}}{\partial t}\frac{dt}{ds} = \mathbf{0}, \tag{6.89}$$

when the initial conditions for the 2D linearised shallow water problem have convergent Fourier series and are continuous at every sample point. Here s is the distance along the path of the characteristic equation and $\mathbf{0} \in \mathbb{R}^3$ denotes the zero vector.

Initially consider the Rossby wave solution, $\mathbf{w}_1 : \mathbb{R} \times \mathbb{R} \times [0, \infty) \to \mathbb{R}^3$ such that,

$$(x, y, t) \mapsto \mathbf{w}_1(x, y, t) := \hat{A} \sum_{p=-\infty}^{\infty} \sum_{q=-\infty}^{\infty} \alpha_{p,q} (\hat{\mathbf{w}}_1)_{p,q} e^{2\pi i (px+qy)}, \tag{6.90}$$

defined as in Equation (6.32) of Section 6.1.2. As the Rossby wave does not change with respect to time, the characteristic equation for the solution is independent of time, so the characteristic speed of the solution is zero. As a result there are no restrictions on the CFL number from this solution.

Next we consider the following inertia-gravity wave, $\mathbf{w}_2 : \mathbb{R} \times \mathbb{R} \times [0, \infty) \to \mathbb{R}^3$ such that,

$$(x, y, t) \mapsto \mathbf{w}_2(x, y, t) := \hat{A} \sum_{p=-\infty}^{\infty} \sum_{q=-\infty}^{\infty} \beta_{p,q} (\hat{\mathbf{w}}_2)_{p,q} e^{2\pi i (px+qy-\hat{\omega}_{p,q} t)}, \tag{6.91}$$

defined as in Equation (6.32) of Section 6.1.2. Substituting (6.91) into (6.89), by the

orthogonality of the Fourier basis functions and considering $\mathbf{w}_2 \neq \mathbf{0}$, we obtain,

$$p\left(\frac{dx}{ds}\right)_{p,q} + q\left(\frac{dy}{ds}\right)_{p,q} - \hat{\omega}_{p,q}\left(\frac{dt}{ds}\right)_{p,q} = 0, \quad \text{for each } p, q \in \mathbb{Z}. \tag{6.92}$$

The index $\cdot_{p,q}$ denotes the relevant variable, for the characteristic equation of the (p,q)th wavenumber component, $p, q \in \mathbb{Z}$. By choosing $\left(\frac{dx}{ds}\right)_{p,q} = \phi p$, $\left(\frac{dy}{ds}\right)_{p,q} = \phi q$ and $\left(\frac{dt}{ds}\right)_{p,q} = \frac{\phi(p^2+q^2)}{\hat{\omega}_{p,q}}$, along with $(t)_{p,q} = 0$ when $s = 0$, we satisfy (6.92) and obtain the characteristic equations,

$$\begin{aligned} &(x(t))_{p,q} = \tfrac{p\hat{\omega}_{p,q}t}{p^2+q^2} + (x_0)_{p,q} \text{ and } (y(t))_{p,q} = \tfrac{q\hat{\omega}_{p,q}t}{p^2+q^2} + (y_0)_{p,q}, && \text{for } p, q \in \mathbb{Z}, \text{ not both zero,} \\ &(x(0))_{p,q} = (x_0)_{p,q} \text{ and } (y(0))_{p,q} = (y_0)_{p,q}, && \text{for } p = q = 0, \end{aligned} \tag{6.93}$$

where $(x(0))_{p,q} = (x_0)_{p,q}$ and $(y(0))_{p,q} = (y_0)_{p,q}$. We obtain the (p,q)th characteristic speeds for the x and y directions, by differentiating the equations for $(x(t))_{p,q}$ and $(y(t))_{p,q}$ in (6.93) respectively, with respect to t for $p, q \in \mathbb{Z}$. Then $\left(\frac{dx}{dt}\right)_{0,0} = \left(\frac{dy}{dt}\right)_{0,0} = 0$ and

$$\left(\frac{dx}{dt}\right)_{p,q} = \frac{p\hat{\omega}_{p,q}}{p^2+q^2} \text{ and } \left(\frac{dy}{dt}\right)_{p,q} = \frac{q\hat{\omega}_{p,q}}{p^2+q^2}, \quad \text{for } p, q \in \mathbb{Z}, \text{ not both zero.} \tag{6.94}$$

This effectively gives each wavenumber component its own CFL number, unlike the linear advection problems we considered previously, where each wavenumber component had the same CFL number. Define the CFL number for the (p,q)th wavenumber component by $h_{p,q} = (h_1)_{p,q} + (h_2)_{p,q}$ such that $h_{0,0} = 0$ and

$$(h_1)_{p,q} := \left|\frac{p\hat{\omega}_{p,q}}{p^2+q^2}\right| \frac{\Delta t}{\Delta x} \text{ and } (h_2)_{p,q} := \left|\frac{q\hat{\omega}_{p,q}}{p^2+q^2}\right| \frac{\Delta t}{\Delta y}, \quad \text{for } p, q \in \mathbb{Z}, \text{ not both zero.} \tag{6.95}$$

The CFL number for the problem is determined by $h = h_1 + h_2$ where,

$$\begin{aligned} h_1 &= \max\left\{ \max_{\substack{p,q \in \mathbb{Z} \\ \text{not both zero}}} \left\{ \left|\frac{p\hat{\omega}_{p,q}}{p^2+q^2}\right| \right\}, 0 \right\} \frac{\Delta t}{\Delta x} \\ \text{and } h_2 &= \max\left\{ \max_{\substack{p,q \in \mathbb{Z} \\ \text{not both zero}}} \left\{ \left|\frac{q\hat{\omega}_{p,q}}{p^2+q^2}\right| \right\}, 0 \right\} \frac{\Delta t}{\Delta y}. \end{aligned}$$

This defines the CFL number using the maximum characteristic speed in each direction for $p, q \in \mathbb{Z}$, ensuring that the domain of dependence of all wavenumber components

of the solution, are contained within the domain of dependence of the finite difference scheme. As $(h_1)_{0,0} = (h_2)_{0,0} = 0$, we maximise the magnitude of the propagation speeds (6.94), over $p, q \in \mathbb{Z}$, not both zero. In the x-direction,

$$\begin{aligned}
\max_{\substack{p,q \in \mathbb{Z} \\ \text{not both zero}}} \left| \left(\frac{dx}{dt} \right)_{p,q} \right| &= \max_{\substack{p,q \in \mathbb{N}_0 \\ \text{not both zero}}} \left(\frac{dx}{dt} \right)_{p,q} \\
&= \max_{p \in \mathbb{N}} \frac{p\hat{\omega}_{p,1}}{p^2+1}, \\
&= \max_{p \in \mathbb{N}} \sqrt{\frac{\phi p^2}{p^2+1} + \frac{f^2}{4\pi^2} \left(\frac{p}{p^2+1} \right)^2}, \\
&\leq \sqrt{\phi + \frac{f^2}{16\pi^2}}. \qquad (6.96)
\end{aligned}$$

The calculations in the y-direction are similar and achieve the same result as in (6.96). Hence the CFL number, for the gravity wave solution in (6.91), is given by $h = h_1 + h_2$ where,

$$h_1 = \sqrt{\phi + \frac{f^2}{16\pi^2}} \frac{\Delta t}{\Delta x} \quad \text{and} \quad h_2 = \sqrt{\phi + \frac{f^2}{16\pi^2}} \frac{\Delta t}{\Delta y}. \qquad (6.97)$$

Performing the same analysis for the remaining gravity wave solution, we obtain the same CFL number for this solution. Therefore the CFL number for the 2D linearised shallow water problem is $h = h_1 + h_2$, where h_1 and h_2 are defined by (6.97). If we consider the case of $f = 0$ for a moment, we find that the propagation speed in h_1 and h_2 becomes $\sqrt{\phi} = \sqrt{gH}$, agreeing with the analysis of Toro [9] for the decoupled one dimensional linearised shallow water system. This makes our CFL number consistent with the 1D decoupled problem.

Now we have determined the CFL number for the 2D linearised shallow water problem, we can use it to help determine if the aliasing errors introduced by the MNIMC scheme for the problem, have a shifted periodic nature. Such a property would allow us to numerically generate perfect observations of the 2D linearised shallow water problem, for use in our strong constraint 4D-Var problem. We investigate the aliasing errors of the MNIMC scheme in the next Section.

6.4.3 Looking for a shifted periodic nature in the MNIMC scheme

When considering the MNIMC scheme for the 1D and 2D linear advection problems, we were able to show that the aliasing errors introduced by the scheme had a shifted periodic nature. We discussed in Section 6.4 that we desire this property in the MNIMC scheme for the 2D linearised shallow water problem, so we can generate perfect observations when the aliasing error in the scheme is periodically zero.

In order to determine if the MNIMC scheme has a shifted periodic nature, we need to define the aliasing error in the MNIMC, when compared to perfect observations of

the system. We define perfect observations similarly to those in Chapters 3 and 5. Let $\tilde{\mathbf{y}}_l \in \mathbb{R}^{3N_xN_y}$ denote a perfect observation of 2D linearised shallow water problem. The vector is structured in the same way as $\mathbf{Z}^n$ in Section 6.1.2 such that,

$$[\tilde{\mathbf{y}}_l]_{(k-1)3N_x+(j-1)3+s} = \begin{cases} u'(x_{j-1}, y_{k-1}, t^l), & \text{for } s = 1, \\ v'(x_{j-1}, y_{k-1}, t^l), & \text{for } s = 2, \\ h'(x_{j-1}, y_{k-1}, t^l), & \text{for } s = 3, \end{cases} \tag{6.98}$$

for $j = 1, \ldots, N_x$ and $k = 1, \ldots, N_y$. Define the vectors $\tilde{\mathbf{u}}_l, \tilde{\mathbf{v}}_l, \tilde{\mathbf{h}}_l \in \mathbb{R}^{N_xN_y}$, for $l \in \mathbb{N}_0$, such that $[\tilde{\mathbf{u}}_l]_{(k-1)N_y+j} = u'(x_{j-1}, y_{k-1}, t^l)$, $[\tilde{\mathbf{v}}_l]_{(k-1)N_y+j} = v'(x_{j-1}, y_{k-1}, t^l)$ and $[\tilde{\mathbf{h}}_l]_{(k-1)N_y+j} = h'(x_{j-1}, y_{k-1}, t^l)$ for $j = 1, \ldots, N_x$ and $k = 1, \ldots, N_y$. Then,

$$\{X^*\tilde{\mathbf{y}}_l\}_{(q-1)3N_x+(p-1)3+r} = \begin{cases} \mathcal{F}_{p,q}(\tilde{\mathbf{u}}_l), & \text{for } r = 1, \\ \mathcal{F}_{p,q}(\tilde{\mathbf{v}}_l), & \text{for } r = 2, \\ \mathcal{F}_{p,q}(\tilde{\mathbf{h}}_l), & \text{for } r = 3. \end{cases} \tag{6.99}$$

As with the previous problems, we define the global error in the MNIMC scheme $\mathbf{r}_l \in \mathbb{R}^{3N_xN_y}$ by,

$$\mathbf{r}_l := \tilde{\mathbf{y}}_l - \tilde{M}^l\tilde{\mathbf{x}}_0, \tag{6.100}$$

where $\mathbf{y}_l := \tilde{\mathbf{y}}_l$ denotes the lth set of perfect observations such that $\{\tilde{\mathbf{y}}_l\}_{(k-1)N_x+j} := w(x_{j-1}, y_{k-1}, l\Delta t)$ for $j = 1, \ldots, N_x$ and $k = 1, \ldots, N_y$. As only aliasing errors are introduced by the MNIMC scheme, $\mathbf{r}_l$ can be interpreted as an additive correction term to correct for aliasing errors in $\tilde{M}^l\tilde{\mathbf{Z}}_0$ such that,

$$\tilde{\mathbf{y}}_l = \tilde{M}^l\tilde{\mathbf{Z}}_0 + \mathbf{r}_l, \tag{6.101}$$

for $l \in \mathbb{N}_0$. We will now attempt to use the same method as in Lemmas 3.12 and 5.6, to determine if the same shifted periodic nature exists in the aliasing error introduced by the MNIMC scheme.

Initially consider the case where $\mathbf{w}_0(x, y)$ is continuous and has a convergent Fourier series. Then applying the 2D DFT implemented by the matrix X^* to $\tilde{\mathbf{y}}_l$

$$\begin{bmatrix} \mathcal{F}_{p,q}(\tilde{\mathbf{u}}_l) \\ \mathcal{F}_{p,q}(\tilde{\mathbf{v}}_l) \\ \mathcal{F}_{p,q}(\tilde{\mathbf{h}}_l) \end{bmatrix} = \sqrt{N_xN_y} \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \hat{A}e^{\hat{D}_{p-1+jN_x, q-1+kN_y} l\Delta t} \hat{A}^{-1} \mathbf{c}_{p-1+jN_x, q-1+kN_y}. \tag{6.102}$$

We want to show that $\mathbf{r}_{nb+d} = \tilde{M}^{nb}\mathbf{r}_d$ for some $b, n, d \in \mathbb{N}_0$. This requires that we identify a value for b where the aliasing error repeats for the first time. In order to determine b, we aim to split $e^{\hat{D}_{p-1+jN_x, q-1+kN_y} l\Delta t}$ into two matrices, one independent of j and k. The matrix independent of j and k will form the matrix $\tilde{M}$ acting on the aliasing error introduced by the scheme at a previous point in time. We could consider

$e^{\hat{D}_{p-1+jN_x,q-1+kN_y}l\Delta t}$ in the form,

$$E_{p-1+jN_x,q-1+kN_y}e^{\Gamma_{p-1+jN_x,q-1+kN_y}l\Delta t}E^*_{p-1+jN_x,q-1+kN_y}, \tag{6.103}$$

as described in (6.42). However, whilst $e^{\Gamma_{p-1+jN_x,q-1+kN_y}l\Delta t}$ may be in a convenient form with exponentials along its main diagonal, $E_{p-1+jN_x,q-1+kN_y}$ is not in a useful form. Therefore, we will consider $e^{\hat{D}_{p-1+jN_x,q-1+kN_y}l\Delta t}$ directly. Consider,

$$e^{\hat{D}_{p-1+jN_x,q-1+kN_y}l\Delta t} = e^{\hat{D}_{p-1+jN_x,q-1+kN_y}[l]_b\Delta t+\hat{D}_{p-1,q-1}(l-[l]_b)\Delta t+\hat{B}_{j,k}(l-[l]_b)\Delta t}, \tag{6.104}$$

the (p,q)th amplification matrix of the MNIMC scheme for $p = 1,\ldots,\frac{N_x+1}{2}$ and $q = 1,\ldots,\frac{N_y+1}{2}$, where $b \in \mathbb{N}_0$ is yet to be defined and

$$\hat{B}_{j,k} = \begin{bmatrix} 0 & 0 & 2\pi jN_x\sqrt{\phi} \\ 0 & 0 & 2\pi kN_y\sqrt{\phi} \\ -2\pi jN_x\sqrt{\phi} & -2\pi kN_y\sqrt{\phi} & 0 \end{bmatrix}. \tag{6.105}$$

Here we can see that we have separated $\hat{D}_{p-1+jN_x,q-1+kN_y}l\Delta t$ into three matrices, one of which is $\hat{D}_{p-1,q-1}(l-[l]_b)\Delta t$. The matrix exponential of this matrix is the (p,q)th matrix of the MNIMC scheme for $p = 1,\ldots,\frac{N_x+1}{2}$ and $q = 1,\ldots,\frac{N_y+1}{2}$. This matrix is independent of j and k, so potentially offers us the opportunity to construct the matrix $\tilde{M}^{l-[l]_b}$, which operates on the aliasing error introduced at a previous point in time. The matrix exponential of $\hat{D}_{p-1+jN_x,q-1+kN_y}[l]_b\Delta t$, would form a part of the aliasing error at a previous point in time. The matrix exponential of $\hat{B}_{j,k}$ would be a remainder term. Separating (6.104) into the multiplication of these matrices would help to demonstrate any shifted periodic nature in the aliasing error introduced by the MNIMC scheme. So what we would like to do is to separate (6.104) into three exponential matrices; $e^{\hat{D}_{p-1+jN_x,q-1+kN_y}[l]_b\Delta t}$, $e^{\hat{D}_{p-1,q-1}(l-[l]_b)\Delta t}$ and $e^{\hat{B}_{j,k}(l-[l]_b)\Delta t}$.

In order to achieve this, we require that the matrices $\hat{D}_{p-1+jN_x,q-1+kN_y}$, $\hat{D}_{p-1,q-1}$ and $\hat{B}_{j,k}$ be commutative [95]. However, each of these matrices is skew-symmetric resulting in $\hat{D}_{p-1,q-1}\hat{B}_{j,k} = (\hat{B}_{j,k}\hat{D}_{p-1,q-1})^T$, so we cannot decompose $e^{\hat{D}_{p-1+jN_x,q-1+kN_y}l\Delta t}$ as we require. If we had been able to decompose $e^{\hat{D}_{p-1+jN_x,q-1+kN_y}l\Delta t}$, we would have required that $e^{\hat{B}_{j,k}(l-[l]_b)\Delta t} = I_3$ for all $j,k \in \mathbb{Z}$, so that the aliasing error is able to become zero periodically. We use this requirement to identify b.

Despite the fact that we are not able to decompose (6.104) into three matrix exponentials, we will continue with our analysis, to see if we could identify a value for b. The aim of this is to understand whether it is solely the formulation of the amplification matrices for the problem as matrix exponentials, that prevents a shifted periodic nature in the aliasing error introduced by the MNIMC scheme, from being identified. The matrix $\hat{B}_{j,k}$, is the matrix $\hat{D}_{jN_x,kN_y}$ under the condition that $f = 0$. Therefore,

$$e^{\hat{B}_{j,k}(l-[l]_b)\Delta t} = E_{jN_x,kN_y}e^{\Gamma_{jN_x,kN_y}(l-[l]_b)\Delta t}E^*_{jN_x.kN_y} \tag{6.106}$$

under the condition that $f = 0$. The eigenvalues of this matrix are,

$$1, \;\; e^{-2\pi i\sqrt{\phi(j^2N_x^2+k^2N_y^2)}(l-[l]_b)\Delta t} \;\text{ and }\; e^{2\pi i\sqrt{\phi(j^2N_x^2+k^2N_y^2)}(l-[l]_b)\Delta t}.$$

Our aim is to find a value for b that will result in $e^{\Gamma_{jN_x,kN_y}(l-[l]_b)\Delta t} = I_3$.

Consider the eigenvalue $e^{-2\pi i\sqrt{\phi(j^2N_x^2+k^2N_y^2)}(l-[l]_b)\Delta t}$. Suppose we re-arrange the CFL number for the problem to obtain,

$$\Delta t = \frac{h}{\sqrt{\phi + \frac{f^2}{16\pi^2}}(N_x + N_y)}. \tag{6.107}$$

Then substituting this into the exponent of the eigenvalue,

$$\frac{-2\pi i\sqrt{\phi(j^2N_x^2 + k^2N_y^2)}(l - [l]_b)h}{\sqrt{\phi + \frac{f^2}{16\pi^2}}(N_x + N_y)}. \tag{6.108}$$

Unfortunately, it is not possible to choose an integer b such that

$$\frac{\sqrt{\phi(j^2N_x^2 + k^2N_y^2)}}{\sqrt{\phi + \frac{f^2}{16\pi^2}}(N_x + N_y)},$$

is an integer for all $j, k \in \mathbb{Z}$. Therefore, even if we could decompose $e^{\hat{D}_{p-1+jN_x,q-1+kN_y}l\Delta t}$, we would not be able to find a shifted periodic nature in the aliasing error introduced by the MNIMC scheme for the 2D linearised shallow water problem. As a result, we have no way to generate perfect observations numerically for our numerical simulations.

The MNIMC scheme for the 2D linearised shallow water problem does not possess a shifted periodic nature in its aliasing error, due to the fact that each wavenumber component of the Fourier series constructing the inertia-gravity wave solutions, have very different phase speeds. The Fourier series for the linear advection problems, had all wavenumber components of the Fourier series solution, travelling through time with the same phase speed. As a result, the wavenumber components all travel together, arriving at each sample point identically to the way they have previously arrived at another sample point. The structure preserving property of the linear advection problem created this property of the solution and hence the shifted periodic nature of the aliasing error introduced by the MNIMC scheme for the problem. This is not possible in the 2D linearised shallow water equations.

The MNIMC scheme for the 2D linearised shallow water problem still provides a theoretical way for defining perfect observations. This can be used to construct a bound for the error in the analysis vector, in the absence of all forms of error, other than those introduced by finite difference approximations in the forward model. This bound would be constructed similarly to those in Lemma 4.4 of Chapter 4 and Lemma 5.11 of

Chapter 5. The bounds derived in Lemmas 5.8-5.10 are required for constructing such a bound. However, as the aliasing error $\mathbf{r}_l$ does not have a shifted periodic nature, this makes constructing a bound on the l_2-norm of the error in the analysis vector slightly harder. Also, since we cannot construct perfect observations numerically, we have no way of testing the ability of the bound to characterise the l_2-norm of the error in the analysis vector. Therefore, we leave the construction of the bound as future work, until a way of constructing perfect observations numerically has been identified. We now end our consideration of the 2D linearised shallow water problem at this point and summarise our findings on the challenges involved in considering this problem.

6.5 Summary

In this Chapter, we have considered the 2D linearised shallow water equations, together with circulant boundary conditions and initial conditions, as our physical system for our the strong constraint 4D-Var data assimilation problem. Choosing this physical system allowed us to extend our analysis of a similar system defined by a single PDE in 2D, to a linear system of PDEs in 2D.

Initially we constructed an analytical solution to the 2D linearised shallow water problem, using a Fourier series solution. This constructed our solution from a Rossby wave and two inertia-gravity wave solutions. We then considered the Upwind and Crank-Nicolson finite difference schemes for solving this problem, constructing a Fourier series representation of the numerical solution. Comparing the Fourier series for the analytical and numerical solutions, found that in order to understand the numerical model error introduced by the considered finite difference schemes, we need to compare the amplification matrices of the solutions. Amplification matrices arise due to the 2D linearised shallow water problem being formed from a system of PDEs. However, we found that we could not compare these matrices by comparing their eigenvalues, as they are not generally simultaneously diagonalisable. Therefore the definitions of numerical dissipation and dispersion defined in Chapter 3 can not be easily applied to this problem.

After considering a strict interpretation of numerical dissipation and dispersion, we settled on the possibility that the polar decomposition of matrices may be a good method for defining the equivalent of numerical dissipation and dispersion, for systems of equations. This involved considering the amplification matrices of the matrix $\tilde{M}^{-1}M$, where $\tilde{M}$ and M are the matrices implementing the MNIMC scheme and the considered imperfect scheme, for the 2D linearised shallow water problem. The amplification matrices are changed into the basis which diagonalises the considered amplification matrix of $\tilde{M}$. Then a polar decomposition is applied. This method requires testing to determine its suitability for determining the equivalent of numerical dissipation and dispersion introduced by finite difference schemes solving systems of PDEs. However, this process would determine the numerically dissipative and dispersive properties of

systems defined by a single PDE, such as the 1D and 2D linear advection problems, considered in Chapters 3 and 4. This makes the approach using the polar decomposition appear plausible.

In order to understand the impact of numerical model error on the analysis vector, as was performed in Chapters 4 and 5, we wanted to consider our strong constraint 4D-Var problem in the absence of all other forms of error. This required the use of perfect observations of the 2D linearised shallow water problem. In this instance, we were unable to make use of the *circshift* function within MATLAB®[74]. Therefore our only option was to construct the MNIMC scheme for the 2D linearised shallow water problem and determine if its aliasing error had a shifted periodic nature. If it did have this property, then it could be used to create perfect observations numerically.

Using the CFL number for the 2D linearised shallow water problem, we found that the aliasing errors introduced by the MNIMC scheme for the problem, did not possess a shifted periodic nature. This is due to the wavenumber components of the gravity wave solutions, all travelling with different phase speeds. Therefore, despite the MNIMC scheme allowing perfect observations to be defined algebraically, it was unable to be used to generate perfect observations numerically.

A bound for the l_2-norm of the error in the analysis vector, due to numerical model error in the considered finite difference schemes, could be constructed using the MNIMC scheme to define perfect observations algebraically. The lack of a shifted periodic nature in the aliasing errors introduced by the MNIMC scheme makes this bound challenging to create. However, without the ability to construct perfect observations numerically, there is no way to determine if the bound is suitable for characterising the error in the analysis vector due to finite difference approximations. The problem of numerically constructing perfect observations of the 2D linearised shallow water problem, needs to be solved before any further analysis is completed.

CHAPTER 7

Conclusions

This thesis has focused on analysing the impact of numerical model error introduced by finite difference schemes, on the accuracy of the analysis vector, formed through strong constraint 4D-Var data assimilation. The accuracy of the analysis vector is relevant to applications which make use of the analysis vector directly and those that make use of it indirectly. Errors enter into the process of strong constraint 4D-Var data assimilation from many different sources. The formulation of the strong constraint 4D-Var cost function aims to minimise the effects of many of these errors. However, the method was developed under the assumption that numerical models for the physical system of interest were perfect, when in reality we know this assumption to be false. Therefore our aim was to quantify and analyse the affects of numerical model error introduced by finite difference schemes on the analysis vector, in order to understand whether the assumption of a perfect model was reasonable and if not, how the effects could be minimised.

Performing such an analysis is important due to the use of strong constraint 4D-Var data assimilation in applications such as numerical weather prediction, where accurate forecasts can influence decisions, possibly saving lives. The physical systems chosen for our investigation were the 1D and 2D linear advection equations and the 2D linearised shallow water equations, each considered together with circulant boundary conditions and initial conditions. These physical systems are used as representative models for more complex systems of interest, as they demonstrate "wave-like flow" [6] present in many meteorologically relevant physical systems. The inclusion of circulant boundary conditions allowed finite difference schemes to be defined for solving the physical systems numerically. This meant that numerical model error was introduced into strong constraint 4D-Var by the use of finite differences to approximate derivatives. Circulant boundary conditions also allowed a spectral approach to be taken with our analysis. By choosing these physical systems, the aim is to demonstrate the relevance of performing such an analysis and present a flavour for the relevant results that can be obtained.

The error introduced by finite difference schemes, for solving the linear advection problems, was described in terms of numerical dissipation and numerical dispersion. Numerical dissipation and dispersion occur when the wavenumber components of the numerical solution are propagated inaccurately, causing the magnitude to artifically decay and the phase to be incorrect respectively. Numerically dissipative and dispersive affects were introduced by the eigenvalues of the matrices implementing the schemes. By examining these eigenvalues, the numerically dissipative and dispersive properties of the schemes could be identified. These properties were determined by the CFL number for the problem. However there was not a method in the literature for measuring the numerically dissipative and dispersive properties of a scheme, to allow a user to choose a CFL number that provided the scheme with the required numerically dissipative and dispersive properties, for a specific task. Therefore dissipative and dispersive metrics were defined to allow a user to gain some understanding of the scale of numerically dissipative and dispersive effects that a scheme introduces, for a given CFL number. Any finite difference scheme for solving the same problem could be used as a reference scheme for use in the metrics, however we advocate the use of the MNIMC scheme as it is numerically non-dissipative and non-dispersive with respect to the resolvable wavenumber components of the numerical solution.

The 2D linearised shallow water problem was formed from a linear system of PDEs. This resulted in the wavenumber components of the numerical solution being propagated by amplification matrices of the scheme. However, there was no extension of numerical dissipation and dispersion to amplification matrices in the literature. Therefore we defined a method for possibly extending the definition through the use of the matrix polar decomposition. This method is able to identify the numerically dissipative and dispersive properties of the resolvable wavenumber components of schemes solving the linear advection problems. It would also classify the MNIMC scheme for the 2D linearised shallow water problem as numerically non-dissipative and non-dispersive with respect to resolvable wavenumber components of the numerical solution. This method shows some promise but requires more rigorous testing.

Initially our investigation considered the effects of finite difference approximations on the analysis vector, in the absence of all other forms of error. This required observations to be taken at every point in space and time of the considered finite difference scheme, the background term to be neglected, the model equations be considered as our physical system and perfect observations be constructed both numerically and algebraically. Constructing perfect observations algebraically was a challenge and would be for any system of PDEs. A finite difference scheme known as the MNIMC scheme was developed to help with this. It allowed perfect observations to be constructed from the state of the system generated by the scheme, together with an additive correction term.

The MNIMC scheme was derived to be numerically non-dissipative and non-dispersive with respect to the resolvable wavenumber components of the numerical solution, for

each physical system. In the case of the linear advection problems, this meant that only aliasing errors in the form of numerical dispersion, were introduced into the numerical solution produced by the scheme. The additive correction term in the perfect observations corrected for this aliasing error. This correction term was found to have a shifted periodic nature for the linear advection problems, due to the shape preserving properties of their analytical solutions. This shifted periodic nature was not present in the correction term for the aliasing error introduced by the MNIMC scheme for the 2D linearised shallow water problem. Despite the MNIMC scheme being computationally expensive, it offered the opportunity to construct perfect observations for the linear advection problems, by running the scheme on a higher resolution grid. More work is required to understand the limitations of this scheme and under what conditions the aliasing error introduced has a shifted periodic nature. However, its high computational expense means that it is a theoretical tool rather than a practical tool for analysing the effects of numerical model error.

Generating perfect observations algebraically using the MNIMC scheme, allowed the analysis vector to be constructed from an amplification matrix acting on a vector containing a discrete sample of the true initial condition, plus an additive correction term. The amplification matrix is solely constructed from the matrices implementing the considered finite difference scheme for the forward model and the MNIMC scheme. This allowed the impact of numerical dissipation and dispersion on resolvable wavenumber components of the discrete sample of the true initial conditions, to be viewed directly. The additive correction term corrected for the aliasing errors introduced by the MNIMC scheme. Analysing the analysis vector in this form revealed,

- increasing the number of sets of observations in the assimilation window does not affect the contribution from wavenumber components propagated by eigenvalues with small magnitudes, when considering numerically dissipative schemes with respect to the resolvable wavenumber components of the numerical solution,

- destructive interference occurs between observations when using numerically non-dissipative and dispersive schemes with respect to the resolvable wavenumber components of the numerical solution, resulting in a loss of information in the analysis vector, which increases as the number of sets of observations in the assimilation window is increased,

- numerically non-dissipative and non-dispersive schemes with respect to the resolvable wavenumber components of the numerical solution, will not necessarily recover the analysis vector due to the effects of aliasing errors.

These results indicate that increasing the number of sets of observations in the assimilation window may not increase the accuracy of the analysis vector, which is a surprising result and needs to be investigated further by re-introducing other forms of error which affect strong constraint 4D-Var data assimilation.

Constructing the analysis vector in this way allows for a deterministic model error operator to be determined for use with weak constraint 4D-Var data assimilation, using the matrices implementing the forward model and MNIMC finite difference schemes. The aliasing errors introduced by the MNIMC scheme are corrected for by using random variables. The shifted periodic nature of the aliasing errors for the linear advection problems means that only a finite number of random variables need be estimated as a part of the weak constraint 4D-Var cost function. This formulation needs testing to determine its ability to reduce the effects of numerical model error from finite difference approximations, on the accuracy of the analysis vector.

Bounds were created in order to understand the effects of finite difference approximations on the l_2-norm of the error in the analysis vector, for the linear advection problems. The bound for each problem depended on the regularity of the initial condition, the numerically dissipative and dispersive properties of the scheme, the number of discretisation points when considering full sets of observations in space and the number of sets of observations in the assimilation window. The use of full sets of observations in space meant that increasing the number of discretisation points resulted in increasing the density of observations in both space and time. The bound for the 1D linear advection problem was found to characterise the worst case behaviour of the error with respect to the number of discretisation points when considering full sets of observations in space and the number of sets of observations in the assimilation window, for numerically dissipative and/or dispersive schemes with respect to the resolvable wavenumber components of the numerical solution. This revealed that:

- The error in the analysis vector remains constant for true initial conditions with resolvable discontinuities, possibly due to Gibb's phenomenon.
- The error in the analysis vector decays at a constant rate as the density of observations in space and time is increased, for initial conditions which appear smooth, with the rate dependent on the regularity of the true initial condition and the numerically dissipative and dispersive properties of the scheme, until a critical regularity is reached where the rate no longer increases with regularity.
- The error in the analysis vector increases as the number of sets of observations in the assimilation window is increased.

These results also indicate that increasing the number of sets of observations in the assimilation window does not increase the accuracy of the analysis vector. However, increasing the density of observations in both space and time appears to improve the accuracy of the analysis vector.

The bound for the l_2-norm of the error in the analysis vector, for the 2D linear advection problem, needs to be developed further to cater for multiplicatively non-separable initial conditions. The bound created in this thesis is for multiplicatively

separable initial conditions only and needs to be analysed further to determine its suitability for characterising the error in the analysis vector. This is motivated by the analysis of the behaviour of some of the summations constructing this bound, which was conducted both analytically and numerically. This revealed that the error in the analysis vector may increase as the density of observations is increased in the x- and y-directions and in time simultaneously, when considering initial conditions containing jump discontinuities. When higher regularity separable initial conditions were considered, the summations indicated a decay in the error as the density of observations was increased in a similar fashion. Results from strong constraint 4D-Var numerical experiments matched the results for initial conditions containing jump discontinuities. The results for initial conditions not containing jump discontinuities appeared to indicate a decay in the l_2-norm of the error in the analysis vector, however more results are required for larger values of N_x and N_y to confirm this behaviour.

Observation errors were re-introduced to the 1D linear advection problem, to understand how finite difference errors and observations errors in the form of white noise, interact to affect the accuracy of the analysis vector. A bound on the l_2-norm of the error in the analysis vector was constructed and compared against numerical results, to understand if the bound could be used to characterise the behaviour of the error in the analysis vector. It was able to characterise the behaviour with respect to the number of discretisation points and the number of sets of observations in the assimilation window. This revealed that there is a critical number of discretisation points when considering full sets of observations where strong constraint 4D-Var data assimilation can be performed, that minimises the impact of finite difference errors from numerically dissipative and/or dispersive finite difference schemes with respect to the resolvable wavenumber components and observation errors. This point is dependent on the variance of the observation errors and the number of sets of observations in the assimilation window.

When considering the numerically non-dissipative and non-dispersive MNIMC scheme, the aliasing errors introduced by the scheme to the analysis vector, were dominated by the observation errors. This resulted in the error in the analysis vector being larger than any other scheme when the number of discretisation points was small and the error always growing so there is no critical number of discretisation points for performing strong constraint 4D-Var data assimilation. It appears from these results that numerical dissipation and/or dispersion is important for reducing the effects of observation errors. Therefore it may be best to perform strong constraint 4D-Var using an imperfect model, provided you have knowledge on the effects of different forms of error.

The contribution to the analysis vector from white noise observation errors was also investigated. It was found that schemes that were numerically dissipative with respect to resolvable wavenumber components, introduced correlations in to the white noise contribution to the analysis vector. This could possibly lead to artifacts in the analysis vector and perhaps its forecast. This is an interesting direction for further research,

and very relevant to the field of inverse problems. Future work is required to analyse the structure of any artifacts and how other forms of error affect them.

The limitations in the results presented in this thesis lie in particular with the consideration of full sets of observations. In reality, it is not possible to take observations of the physical system at every point in space and time of the physical system. By taking observations at every point in space and time, this has resulted in the number of sets of observations in the assimilation window of the physical system increasing as either the number of discretisation points in space (N) and the number of sets of observations in the assimilation window (L) are increased. This has resulted in this thesis investigating the impact of changing the density of observations in space and time by changing N and the number of sets of observations in the assimilation window by altering L. It is important to investigate how increasing the number of discretisation points impacts the contribution of model errors to the analysis vector. This result would be more practically relevant for implementing 4D-Var data assimilation. This can be achieved by performing a similar analysis to that conducted in this thesis, but fixing the observation points whilst increasing the number of discretisation points of the numerical model. This could possibly result in the problem becoming ill-posed, requiring the background term to be introduced to the analysis, adding background errors to the problem.

Another limitation associated with the results of this thesis is the assumption that the cost function is exactly minimised numerically. The problem considered was numerically solved using the pre-conditioned conjugate gradient (PCG) method and was able to converge to the exact solution obtained theoretically. The bounds derived in this thesis are able to represent the behaviour of the error in the analysis vector as a consequence of this. Future work needs to take into account the fact that the numerical method may not be able to achieve the theoretical analysis vector. This could be achieved by considering the solution path of the numerical method when calculating the analysis vector.

We end this Chapter with a summary of the key results discussed, that have arisen from the research conducted in this thesis:

- Increasing the number of sets of observations in the assimilation window does not necessarily improve the accuracy of the analysis vector.
- There is an optimal number of discretisation points when considering full sets of observations, where the effects of numerical model error and observation errors on the analysis vector are minimised for the 1D linear advection problem, for numerically dissipative and/or dispersive schemes with respect to the resolvable wavenumber components of the numerical solution.
- White noise observations can become correlated in the analysis vector when numerically dissipative schemes are considered, possibly leading to artifacts in the analysis vector.

- Numerically non-dissipative and dispersive schemes with respect to the resolvable wavenumber components, introduce destructive interference between wavenumber components resulting in a loss of information in the analysis vector. However they do not introduce correlations into the analysis vector from white noise observations.
- Numerical dissipation and/or dispersion in resolvable wavenumber components of the numerical solution, helps to counter the effects of observation errors, so can be an advantage.
- A numerically non-dissipative and non-dispersive finite difference scheme with respect to the resolvable wavenumber components of the numerical solution, was developed.
- The creation of numerically dissipative and dispersive metrics for determining the numerically dissipative and dispersive properties of a scheme for solving the 1D or 2D linear advection problems, for a range of CFL numbers.

Finally we suggest possible directions for future research:

- Perform a similar analysis to investigate the affects from other sources of error in strong constraint 4D-Var data assimilation, on the accuracy of the analysis vector, especially the effects introduced by the sparsity and spread of observations.
- Develop a definition for numerical dissipation and dispersion introduced by amplification matrices, using the matrix polar decomposition method.
- Refine the bound for the l_2-norm of the error in the analysis vector for the 1D linear advection problem, so it reflect the behaviour of the error for the MNIMC scheme.
- Investigate the affect of the proposed deterministic model error operator for use with weak constraint 4D-Var data assimilation of the 1D linear advection problem.
- Construct a bound for the l_2-norm of the error in the analysis vector in terms of Gibb's phenomenon, to determine how this affects the analysis vector for true initial conditions with resolvable discontinuities.
- To understand the conditions under which the aliasing error introduced by the MNIMC scheme, has a shifted periodic nature.
- Investigate the behaviour of the l_2-norm of the error in the analysis vector,for the 2D linear advection problem, when considering initial conditions containing resolvable discontinuities.

APPENDIX A

Corrections to the 1D Bounds

This chapter of the appendix derives corrections to the statements and proofs for the main bounds utilised in Chapter 4. The considered bounds are the bound on the 1D Fourier coefficients, explored in Section A.1 and the bound on the error in the 1D DFT, considered in Section A.2. The Lemmas in this Appendix are taken from their cited source, with some of the notation adapted to reflect the variables considered in this thesis in order to avoid confusion.

A.1 The bound on the 1D Fourier coefficients

Initially we consider the bound on the 1D Fourier coefficients. Consider the Fourier series for a 2π-periodic function $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto f(x)$, in sine and cosine form,

$$f(x) = \sum_{k=0}^{\infty} \{\alpha_k \cos(kx) + \beta_k \sin(kx)\}. \tag{A.1}$$

Carslaw [61, p. 269] provides a bound for the coefficients α_k and β_k in (A.1) along with an outline for the proof of this bound. The statement for this bound is reproduced from Carslaw [61] in Lemma A.1.

Lemma A.1. *If the function $f(x)$ and its differential coefficients, up to the $(r-1)$th, are bounded and continuous and otherwise satisfy Dirichlet's Conditions in the interval $-\pi < x < \pi$, and*

$$f^{(s)}(-\pi + 0) = f^{(s)}(\pi - 0),$$

$s = 0, 1, \ldots, (r-1)$ and if the rth differential coefficient is bounded and otherwise satisfies Dirichlet's Conditions in the same interval, the coefficients in the Fourier's Series for $f(x)$ will be less in absolute value than $\frac{\gamma}{k^{r+1}}$, where γ is some positive number

independent of k.

In other words under these conditions,

$$|\alpha_k| < \frac{\gamma}{k^{r+1}}, \text{ and } |\beta_k| < \frac{\gamma}{k^{r+1}}. \tag{A.2}$$

The statement refers to Dirichlet's conditions which can be found in [61, 62]. These are the sufficient conditions for the Fourier series of $f(x)$ to be convergent [62]. Consider the statement for when $k = 0$. No exception is made for this case in Lemma A.1, where it appears that the denominator of the bounds in (A.2) are equal to zero. Let $f_k \in \mathbb{C}$ denote the kth Fourier coefficient for the exponential form of the Fourier series for $f(x)$. When $k \in \mathbb{Z} \setminus \{0\}$, the bounds in (A.2) can be used to form a bound on f_k using [60],

$$f_k = \begin{cases} \frac{\alpha_k - i\beta_k}{2}, & \text{for } k > 0, \\ \frac{\alpha_0}{2}, & \text{for } k = 0, \\ \frac{\alpha_{-k} + i\beta_{-k}}{2}, & \text{for } k < 0. \end{cases} \tag{A.3}$$

A bound on the coefficients of the Fourier series in exponential form is provided by Henson [66, Theorem 3.5 p. 48] without proof and by Briggs and Henson [60, Theorem 6.2, p. 187] together with an outline of the proof. The statements of Henson [66] and Briggs and Henson [60] define $f(x)$ to be a $2A$-periodic function and consider $f(x)$ over the domain $(-A, A)$. The coefficients f_k, $k \in \mathbb{Z}$ are the Fourier coefficients for the exponential form of the Fourier series of $f(x)$ over this domain. The statement for this bound in Lemma A.2 is taken from Henson [66].

Lemma A.2. *Let $f(x)$ be bounded and satisfy Dirichlet's conditions in $(-A, A)$. Then the Fourier coefficients for the $2A$-periodic extension of $f(x)$ satisfy,*

$$|f_k| \leq \frac{M}{|k|}, \tag{A.4}$$

for some constant $M > 0$ independent of k.

For any $r \geq 1$, assume that $f(x)$ and it's derivatives, up to the $(r-1)^{st}$, are bounded, continuous, and satisfy Dirichlet's conditions in $(-A, A)$. For $s = 0, 1, \ldots, r-1$, assume that,

$$\lim_{x \to -A^-} f^{(s)}(x) = \lim_{x \to A^+} f^{(s)}(x).$$

If the rth derivative is bounded and satisfies Dirichlet's conditions in the same interval, then the Fourier coefficients for the $2A$-periodic extension of $f(x)$ satisfy,

$$|f_k| \leq \frac{C}{|k|^{r+1}} \tag{A.5}$$

for some constant $C > 0$ independent of k.

Here $f^{(s)}(x)$ denotes the sth derivative of $f(x)$ with respect to x. The statement in Lemma A.2 also makes no exception for the case of $k = 0$. In order to verify the result of Lemma A.2 for $k \in \mathbb{Z} \setminus \{0\}$, investigate the case of $k = 0$ and identify the constants M and C in (A.4) and (A.5) respectively, we follow the directions of Carslaw [61] and Briggs and Henson [60] and construct a proof for Lemma A.2. This proof requires a statement for the *2nd Mean Value Theorem (MVT) for Integrals*. This is provided in Lemma A.3, taken from Courant [86].

Lemma A.3. *Let us suppose that the function $f(x)$ is monotonic and continuous in the interval $a \leq x \leq b$, and that the derivative $f'(x)$ is continuous; and let us further suppose that $\phi(x)$ is an arbitrary function continuous in the same interval. This requires the second mean value theorem of the integral of calculus is expressed as follows. There exists ξ, such that $a \leq \xi \leq b$, for which*

$$\int_a^b f(x)\phi(x)\mathrm{d}x = f(a)\int_a^\xi \phi(x)\mathrm{d}x + f(b)\int_\xi^b \phi(x)\mathrm{d}x.$$

Now we have a statement for the 2nd MVT for Integrals, we embark upon a proof for a bound on the coefficients of the 1D Fourier series in exponential form. This is found in Lemma A.4 where it should be noted that the statement of Lemma A.2 has been modified to use the definition of regularity in Definition 3.8. Dirichlet's conditions do not form a part of the requirements of Lemma A.2, as the conditions satisfy Theorem 3.1, ensuring that the considered function and its derivatives have a convergent Fourier series. The function $f^{(\alpha)}(x)$ is not required to be bounded over $(0,T)$ for $\alpha = 0,\ldots,r-1$, as $f^{(\alpha)}(x)$ is bounded over $(0,T)$ as a consequence of its continuity. The statement of Lemma A.4 is also found in Lemma 4.3 of Chapter 4.

Lemma A.4. *Let $r \in \mathbb{N}_0$ denote the maximum number of times the function $f : (0,T) \to \mathbb{R}$, $x \mapsto f(x)$ can be differentiated such that $f^{(\alpha)}(x)$ is continuous and piecewise differentiable over $(0,T)$, for $\alpha = 0,\ldots,r-1$ and $f^{(r)}(x)$ is piecewise continuous over $(0,T)$ ie: $f(x)$ has regularity r over $(0,T)$. Also let,*

$$\lim_{x\to 0^+} f^{(\alpha)}(x) = \lim_{x\to T^-} f^{(\alpha)}(x), \tag{A.6}$$

for $\alpha = 0,\ldots,r-1$ and $f(x)$ be piecewise monotone over $(0,T)$ and $f^{(r)}(x)$ be bounded

and piecewise monotone over $(0,T)$*. Then the coefficients of the Fourier series for* $f(x)$*, given by* $f_k \in \mathbb{C}$*,* $k \in \mathbb{Z}$*, can be bounded such that,*

$$|f_k| \leq \begin{cases} D_1, & \text{for } k = 0, \\ \frac{D_2}{|k|^{r+1}}, & \text{for } k \in \mathbb{Z} \setminus \{0\}, \end{cases} \tag{A.7}$$

where D_1 *and* D_2 *are constants independent of* k*.*

Proof. The kth coefficient of the Fourier series for $f(x)$ over $(0,T)$, $T > 0$ is given by,

$$f_k = \frac{1}{T} \int_0^T f(x) e^{\frac{-2\pi i k x}{T}} \mathrm{d}x. \tag{A.8}$$

When $r \geq 1$, $f^{(\alpha)}(x)$ is continuous and piecewise differentiable with respect to x, for $\alpha = 0, \ldots, r-1$. This allows integration by parts to be applied to f_k, r-times, using (A.6) for $\alpha = 0, \ldots, r-1$,

$$f_k = \begin{cases} \frac{1}{T} \int_0^T f(x) \mathrm{d}x, & \text{for } k = 0, \\ \frac{1}{T} \left(\frac{-iT}{2\pi k}\right)^r \int_0^T f^{(r)}(x) e^{\frac{-2\pi i k x}{T}} \mathrm{d}x, & \text{for } k \in \mathbb{Z} \setminus \{0\}. \end{cases} \tag{A.9}$$

Combining (A.9) with the case of $r = 0$, results in (A.9) holding true for all r.

As $f(x)$ and $f^{(r)}(x)$ are piecewise monotone over $(0,T)$, their domain can be broken up into a finite number of open partial intervals where the respective function is continuous and monotonic. For $f(x)$ define the partial intervals, (a_j, a_{j+1}), $j = 1, \ldots, s_1$ such that $0 = a_1 < a_2 < \ldots < a_{s_1} < a_{s_1+1} = T$, for some $s_1 \in \mathbb{N}$. For $f^{(r)}(x)$ define the partial interval (b_j, b_{j+1}), $j = 1, \ldots, s_2$ such that $0 = b_1 < b_2 < \ldots < b_{s_2} < b_{s_2+1} = T$ for some $s_2 \in \mathbb{N}$. If $r = 0$, $a_j = b_j$ for all $j = 1, \ldots, s_1$, and $s_1 = s_2$. Then,

$$f_k = \begin{cases} \frac{1}{T} \sum_{j=1}^{s_1} \int_{a_j}^{a_{j+1}} f(x) \mathrm{d}x, & \text{for } k = 0, \\ \frac{1}{T} \left(\frac{-iT}{2\pi k}\right)^r \sum_{j=1}^{s_2} \int_{b_j}^{b_{j+1}} f^{(r)}(x) e^{\frac{-2\pi i k x}{T}} \mathrm{d}x, & \text{for } k \in \mathbb{Z} \setminus \{0\}. \end{cases} \tag{A.10}$$

As $f(x)$ is continuous and monotonic over each (a_j, a_{j+1}), the 2nd Mean Value Theorem for integrals can be applied to the case of $k = 0$ in (A.10). This implies that there exists $\xi_j \in (a_j, a_{j+1})$ such that,

$$\int_{a_j}^{a_{j+1}} f(x) \mathrm{d}x = \left\{ \left(\lim_{x \to a_j^+} f(x) \right) \int_{a_j}^{\xi_j} 1 \mathrm{d}x + \left(\lim_{x \to a_{j+1}^-} f(x) \right) \int_{\xi_j}^{a_{j+1}} 1 \mathrm{d}x \right\}. \tag{A.11}$$

for $j = 1, \ldots, s_1$. Similarly, as $f^{(r)}(x)$ is continuous and monotonic and $e^{\frac{-2\pi i k x}{T}}$ is continuous over each (b_j, b_{j+1}), applying the 2nd Mean Value Theorem to the case of

$k \in \mathbb{Z} \setminus \{0\}$ in (A.10) implies that there exists $\psi_j \in (b_j, b_{j+1})$ such that,

$$\int_{b_j}^{b_{j+1}} f^{(r)}(x) e^{\frac{-2\pi i k x}{T}} \mathrm{d}x = \left\{ \lim_{x \to b_j^+} f^{(r)}(x) \int_{b_j}^{\psi_j} e^{\frac{-2\pi i k x}{T}} \mathrm{d}x + \lim_{x \to b_{j+1}^-} f^{(r)}(x) \int_{\psi_j}^{b_{j+1}} e^{\frac{-2\pi i k x}{T}} \mathrm{d}x \right\}, \tag{A.12}$$

for $j = 1, \ldots, s_2$.

When $r \in \mathbb{N}$, $f(x)$ is continuous which implies that $f(x)$ is bounded. Let $v_1, v_2 \in \mathbb{R}$ denote the bounds for $f(x)$ and $f^{(r)}(x)$ in the interval $(0, T)$, respectively. When $r = 0$, $v_1 = v_2$. Initially, consider the case of $k = 0$. Using (A.10) and (A.11),

$$f_0 = \frac{1}{T} \sum_{j=1}^{s_1} \left\{ \left(\lim_{x \to a_j^+} f(x) \right) (\xi_j - a_j) + \left(\lim_{x \to a_{j+1}^-} f(x) \right) (a_{j+1} - \xi_j) \right\}. \tag{A.13}$$

We now find the magnitude of f_0 by applying the triangle inequality,

$$\begin{aligned} |f_0| &\leq \frac{1}{T} \sum_{j=1}^{s_1} \left\{ \left| \lim_{x \to a_j^+} f(x) \right| |\xi_j - a_j| + \left| \lim_{x \to a_{j+1}^-} f(x) \right| |a_{j+1} - \xi_j| \right\} \\ &\leq \frac{v_1}{T} \sum_{j=1}^{s_1} \left(|a_{j+1} - \xi_j| + |\xi_j - a_j| \right), \text{ as } |f(x)| \leq v_1 \text{ over } (0, T), \\ &= v_1. \end{aligned} \tag{A.14}$$

Now consider the case of $k \in \mathbb{Z} \setminus \{0\}$ using (A.10) and (A.12),

$$\begin{aligned} f_k &= \frac{1}{T} \left(\frac{-iT}{2\pi k} \right)^r \left(\frac{iT}{2\pi k} \right) \sum_{j=1}^{s_2} \left\{ \left(\lim_{x \to b_j^+} f^{(r)}(x) \right) \left(e^{\frac{-2\pi i k \xi_j}{T}} - e^{\frac{-2\pi i k a_j}{T}} \right) \right. \\ &\quad \left. + \left(\lim_{x \to b_{j+1}^-} f^{(r)}(x) \right) \left(e^{\frac{-2\pi i k a_{j+1}}{T}} - e^{\frac{-2\pi i k \xi_j}{T}} \right) \right\}. \end{aligned} \tag{A.15}$$

Finding the magnitude of f_k in (A.15) and applying the triangle inequality results in,

$$\begin{aligned} |f_k| &\leq \frac{1}{T} \left(\frac{T^{r+1}}{(2\pi|k|)^{r+1}} \right) \sum_{j=1}^{s_2} \left\{ \left| \lim_{x \to b_j^+} f^{(r)}(x) \right| \left| e^{-2\pi i k \psi_j} - e^{-2\pi i k b_j} \right| \right. \\ &\quad \left. + \left| \lim_{x \to b_{j+1}^-} f^{(r)}(x) \right| \left| e^{-2\pi i k b_{j+1}} - e^{-2\pi i k \psi_j} \right| \right\} \\ &\leq \frac{4 v_2 s_2 T^r}{(2\pi|k|)^{r+1}}, \text{ as } |f^{(r)}(x)| \leq v_2 \text{ over } (0, T). \end{aligned} \tag{A.16}$$

Let $D_1 = v_1$ and $D_2 = \frac{4 v_2 s_2 T^r}{(2\pi)^{r+1}}$. Therefore, by (A.14) and (A.16),

$$|f_k| \leq \begin{cases} D_1, & \text{for } k = 0, \\ \frac{D_2}{|k|^{r+1}}, & \text{for } k \in \mathbb{Z} \setminus \{0\}. \end{cases} \tag{A.17}$$

The constants D_1 and D_2 are independent of k. However it should be noted that they are dependent upon the regularity of $f(x)$ over $(0,T)$ ie: r. □

A.2 The bound on the error in the 1D DFT

We now consider the bound on the error in the coefficient identified using the 1D DFT, when compared to the coefficient of the Fourier series for the same function, for the same resolvable wavenumber component. Briggs and Henson [60, Theorem 6.3 p.188] provides a statement for this bound, for the $2A$-periodic function $f(x)$ considered in Lemma A.2, along with a sketch proof. The statement of this bound is cited as coming from Henson [66]. Examining Henson [66], the bound is defined in the form of Theorems 3.6 and 3.7 for the same function with regularities $r \in \mathbb{N}$ and $r = 0$, respectively. However, the proof of Theorem 3.7 contains two small mistakes which need to be corrected. The statement of Theorem 3.7 in [66] is stated in Lemma A.5

Lemma A.5. *Assume $f(x) = 0$ for $|x| \geq A$ and that f is bounded in $(-A, A)$ and continuous except for a finite number, q, of jump discontinuities. Suppose also that $f(x)$ and its first derivative satisfy Dirichlet's conditions on every open sub-interval of $(-A, A)$ that does not contain a jump discontinuity of f. Let the Fourier transform values $\hat{f}(\omega_k)$ be approximated by,*

$$\hat{f}_k = \sqrt{\frac{2}{\pi}} \frac{A}{N_x} \sum_{n=-\frac{N_x}{2}+1}^{\frac{N_x}{2}} f(x_n) e^{\frac{-2\pi i n k}{N_x}}, \tag{A.18}$$

for $k = -\frac{N_x}{2} + 1, -\frac{N_x}{2} + 2, \ldots, \frac{N_x}{2}$. Then the error in this approximation satisfies

$$\left| \hat{f}(\omega_k) - \hat{f}_k \right| \leq \frac{D_3}{N_x}, \tag{A.19}$$

where C is a constant that is independent of k or N_x.

The result of this proof can be re-written using the notation of this thesis, to create,

$$\begin{array}{rcll} \left| \frac{1}{\sqrt{N_x}} \mathcal{F}_k(f) - f_{k-1} \right| & \leq & \frac{D_3}{N_x}, & \text{for } k = 1, \ldots, \lfloor \frac{N_x}{2} \rfloor + 1, \\ \left| \frac{1}{\sqrt{N_x}} \mathcal{F}_k(f) - f_{N_x - k - 1} \right| & \leq & \frac{D_3}{N_x}, & \text{for } k = \lfloor \frac{N_x}{2} \rfloor + 2, \ldots, N_x. \end{array} \tag{A.20}$$

Here $f_k \in \mathbb{C}$ for $k \in \mathbb{Z}$, denotes the Fourier coefficient for the kth Fourier basis function of $f(x)$.

The proof of Theorem 3.7 in [66] will be corrected in the following. This does not affect the overall result of the Theorem, but does alter the coefficient in the bound. The proof of Theorem 3.7 ($r = 0$) in Henson [66] is different to that of Theorem 3.6 ($r \in \mathbb{N}$) in Henson [66]. This is due to the proof for $r \in \mathbb{N}$ relying upon the Poisson summation representation of the 1D DFT coefficient as in equation (3.22) of Section 3.4.1. The Poisson summation requires that the Fourier series for the function be equal to the function at every sample point of the domain. In the instance of $r = 0$, this may not be true as the function possesses a finite number of discontinuities.

In order to make use of the Poisson summation, the proof uses the non-uniqueness of the 1D DFT of a function. A new function $g : \mathbb{R} \to \mathbb{R}$ such that $x \mapsto g(x)$ is defined. The aim of this function is for it to have the same 1D DFT as $f(x)$ and have regularity $r = 1$. Therefore $g(x)$ is defined by linearly interpolating $f(x)$ across its discontinuities. The Fourier coefficient for $g(x)$, denoted by $g_k \in \mathbb{C}$ for all $k \in \mathbb{Z}$, can be used as follows,

$$\begin{aligned} \left| \frac{1}{\sqrt{N_x}} \mathcal{F}_k(f) - f_{k-1} \right| &= \left| \frac{1}{\sqrt{N_x}} \mathcal{F}_k(f) - g_{k-1} + g_{k-1} - f_{k-1} \right|, \\ &\leq \left| \frac{1}{\sqrt{N_x}} \mathcal{F}_k(f) - g_{k-1} \right| + |g_{k-1} - f_{k-1}|, \end{aligned} \tag{A.21}$$

for $k = 1, \ldots, \left\lfloor \frac{N_x}{2} \right\rfloor + 1$ and,

$$\begin{aligned} \left| \frac{1}{\sqrt{N_x}} \mathcal{F}_k(f) - f_{N_x-k-1} \right| &= \left| \frac{1}{\sqrt{N_x}} \mathcal{F}_k(f) - g_{N_x-k-1} + g_{N_x-k-1} - f_{N_x-k-1} \right|, \\ &\leq |\mathcal{F}_k(f) - g_{N_x-k-1}| + |g_{N_x-k-1} - f_{N_x-k-1}|, \end{aligned} \tag{A.22}$$

for $k = \left\lfloor \frac{N_x}{2} \right\rfloor + 2, \ldots, N_x$. We can then bound $|\mathcal{F}_k(f) - g_{k-1}|$ and $|\mathcal{F}_{N_x-k-1}(f) - g_{N_x-k-1}|$ using Theorem 3.6 in Henson [66], stated in Lemma 4.3, for $r = 1$.

We consider the function $f(x)$ defined in Lemma A.4 with $r = 0$. The function $g(x)$ defined by Henson [66], is created from the function $f(x)$ using the following sets [66],

$$\hat{X}_T = \{\hat{x} \in [0, T) | f(x) \text{ is a jump discontinuity}\}, \tag{A.23}$$

$$\hat{Z} = \left\{ (\hat{y}, \hat{z}) \in [0, 1) \times [0, 1) | \hat{y} = \left\lfloor \frac{\hat{x}}{\Delta x} \right\rfloor \Delta x, \;\; \hat{z} = \left\lceil \frac{\hat{x}}{\Delta x} \right\rceil \Delta x, \;\; \hat{x} \in \hat{X}_T \right\} \tag{A.24}$$

The set $\hat{Z}$ is then the ordered pairs $(\hat{y}, \hat{z})$, such that the function $f(x)$ contains a jump discontinuity over the sub-domain $(\hat{y}, \hat{z})$. Let $w = |\hat{Z}|$ and give each element of $\hat{Z}$ an index $j = 1, \ldots, w$. Also, let $\hat{z}_0 = 0$ and $\hat{y}_{w+1} = T$. Then $g(x)$ is defined such that $g : \mathbb{R} \to \mathbb{R}$,

$$x \mapsto g(x) = \begin{cases} f(x), & \text{for } x \in \cup_{j=0}^{Q} [\hat{z}_j, \hat{y}_{j+1}], \\ \frac{[f(\hat{z}_j) - f(\hat{y}_j)]}{\Delta x}(x - \hat{y}_j) + f(\hat{y}_j), & \text{for } x \in [\hat{y}_j, \hat{z}_j], \end{cases} \tag{A.25}$$

and $g(x) = g(x + 1)$. If the subdomain (x_q, x_{q+1}) contains a point of discontinuity for some $q = 0, \ldots, N_x$, then $g(x)$ possesses a linear interpolation across the domain,

between points $(\hat{y}_j, f(\hat{y}_j))$ and $(\hat{z}_{j+1}, f(\hat{z}_{j+1}))$, for some $j = 1, \ldots, w$. If $(\hat{x}_q, \hat{x}_{q+1})$ does not possess a point of discontinuity, then $g(x)$ is equal to $f(x)$ across the sub-domain. However, suppose that N_x is sufficiently large and that there only exists a discontinuity at $x = \Delta x$. Then $g(\Delta x) = f(\Delta x)$ by (A.25). The result is that $g(x) = f(x)$ so $g(x)$ is not continuous at $x = \Delta x$, resulting in a regularity of $r = 0$, preventing the Poisson summation from being used to represent the coefficient found by the 1D DFT.

In order to tackle this problem and identify the coefficients of the bound, we create a new function which has the same 1D DFT as that possessed by $f(x)$ and regularity $r = 1$. Define the set,

$$Q = \left\{ \hat{x} \in \hat{X}_T \,|\, \hat{x} = s\Delta x, s \in \mathbb{Z} \right\}. \tag{A.26}$$

This is the set of discontinuities in $f(x)$ which coincide with a grid point in space. Now we define subsets of this set,

$$Q_1 = \left\{ \left[\hat{x} - \frac{\Delta x}{2} \right]_T \,|\, \hat{x} \in Q \right\}, \quad Q_2 = \left\{ \left[\hat{x} + \frac{\Delta x}{2} \right]_T \,|\, \hat{x} \in Q \right\}, \tag{A.27}$$

where $[\cdot]_T$ denotes modulo T. Now we can re-define the sets $\hat{Y}$ and $\hat{Z}$,

$$\hat{Y} = \left\{ \hat{y}_j \in [0, T) \,|\, \hat{y}_j = \left\lfloor \frac{\hat{x}_j}{\Delta x} \right\rfloor \Delta x, \;\; \hat{x}_j \in \hat{X} \cup Q_1 \cup Q_2 \right\}, \tag{A.28}$$

$$\hat{Z} = \left\{ \hat{z}_j \in [0, T) \,|\, \hat{z}_j = \left\lceil \frac{\hat{x}_j}{\Delta x} \right\rceil \Delta x, \;\; \hat{x}_j \in \hat{X} \cup Q_1 \cup Q_2 \right\}. \tag{A.29}$$

This ensures that given any $\hat{x} \in Q$ a linear interpolation over $f(x)$ is created over $(\hat{x} - \Delta x, \hat{x})$ and $(\hat{x}, \hat{x} + \Delta x)$. This ensures that there is no longer a discontinuity present in $g(x)$ at $\hat{x} \in Q$.

We now continue with the bound on (A.21) and (A.22). Following Henson [66] and using the amended $g(x)$,

$$\begin{aligned} |\mathcal{F}_k(\tilde{\mathbf{x}}_0) - f_{k-1}| &\leq |\mathcal{F}_k(\tilde{\mathbf{x}}_0) - g_{k-1}| + |g_{k-1} - f_{k-1}|, \\ &\leq \frac{D_3}{N_x^2} + |g_{k-1} - f_{k-1}|, \end{aligned}$$

for $k = 1, \ldots, \left\lfloor \frac{N_x}{2} \right\rfloor + 1$ and

$$\begin{aligned} |\mathcal{F}_k(\tilde{\mathbf{x}}_0) - f_{N_x-k-1}| &\leq |\mathcal{F}_k(\tilde{\mathbf{x}}_0) - g_{N_x-k-1}| + |g_{N_x-k-1} - f_{N_x-k-1}|, \\ &\leq \frac{D_3}{N_x^2} + |g_{N_x-k-1} - f_{N_x-k-1}|, \end{aligned}$$

for $k = \left\lfloor \frac{N_x}{2} \right\rfloor + 2, \ldots, N_x$, by Lemma 4.3 for $r = 1$.

For some $k \in \mathbb{Z}$, we have that,

$$\begin{aligned}
|g_k - f_k| &= \left| \int_0^T [g(x) - f(x)] e^{-2\pi i k x} \mathrm{d}x \right|, \\
&= \left| \sum_{j=1}^{w} \int_{\hat{y}_j}^{\hat{x}_j} [g(x) - f(x)] e^{-2\pi i k x} \mathrm{d}x \right|, \\
&= \sum_{j=1}^{w} \int_{\hat{y}_j}^{\hat{z}_j} |g(x) - f(x)| \, \mathrm{d}x, \\
&\quad \text{by the generalised fundamental theorem of calculus (FTC),} \\
&\leq 2 v_1 w \Delta x,
\end{aligned}$$

where $w = |\hat{Y}| = |\hat{Z}|$ and v_1 is the bound on the function $f(x)$, so consequently bounds $g(x)$. The factor w is missing from the proof by Henson [66], probably due to a typographical error. As $\Delta x = \frac{1}{N_x}$, we then have that,

$$\begin{aligned}
|\mathcal{F}_k(\tilde{\mathbf{x}}_0) - f_{k-1}(x)| &\leq \frac{D_2[4 + 2\zeta(2)]}{N_x^2} + \frac{2 v_1 w}{N_x}, \\
&\leq \frac{D_2[4 + 2\zeta(2)] + 2 v_1 w}{N_x},
\end{aligned} \tag{A.30}$$

for $k = 1, \ldots, \left\lfloor \frac{N_x}{2} \right\rfloor + 1$ and

$$\begin{aligned}
|\mathcal{F}_k(\tilde{\mathbf{x}}_0) - f_{N_x - k - 1}(x)| &\leq \frac{D_2[4 + 2\zeta(2)]}{N_x^2} + \frac{2 v_1 w}{N_x}, \\
&\leq \frac{D_2[4 + 2\zeta(2)] + 2 v_1 w}{N_x}.
\end{aligned} \tag{A.31}$$

Hence when $r = 0$, $D_3 = \frac{D_2[4+2\zeta(2)]+2v_1 w}{N_x}$ as stated in Lemma 4.3.

APPENDIX B

Numerical Orders of Convergence for the 1D Linear Advection Problem

In this chapter of the Appendix, we consider the numerical results for the 1D linear advection problem. Consider the summations S_1 to S_6 of Section 4.39. Let E_k denote the magnitude of the error in S_k when compared to zero, for $k = 1, \ldots, 6$. Assume the error has the following form,

$$E_k \approx C_k N_x^{\alpha_k} L^{\beta_k}. \tag{B.1}$$

Given this, the order of convergence with respect to both N_x and L can be calculated. In order to identify α_k, N_x is varied whilst L remains constant. We require N_x to be odd due to the use of the MNIMC scheme, so N_x is chosen such that $N_x = 3^\gamma$ for $\gamma = 2, \ldots, 7$ for the Upwind and Preissman Box schemes and $\gamma = 2, \ldots, 12$ for the Lax-Wendroff scheme, whilst $L = 4$. Let E_{k,N_x} denote the magnitude of the error between S_k and zero, as $N_x \to \infty$. Then,

$$\log\left(\frac{E_{k,3N_x}}{E_{k,N_x}}\right) / \log(3) = \log\left(\frac{C_k(3N_x)^{\alpha_k}}{C_k N_x^{\alpha_k}}\right) / \log(3) = \alpha_k, \tag{B.2}$$

calculates α_k the order of convergence of E_k with respect to N_x. A similar calculation can be performed to identify β_k. In this instance, L is chosen such that $L = 2^\delta$ for $\delta = 0, \ldots, 9$, whilst $N_x = 3^7$. Let $E_{k,L}$ denote the magnitude of the error in S_k when compared to zero as $L \to \infty$. The Tables in the following Sections present the numerical values for α_k and β_k for $k = 1, \ldots, 6$.

Consider the approximations of $|1 - \nu_p|$ and ξ_p for the Upwind scheme, in (4.38) and (4.40) respectively, for $p = 2, \ldots, \frac{N_x+1}{2}$. These can be used to try and determine analytical orders of convergence to zero for each S_k, for the Upwind scheme. As (4.38)

identifies that $|1-\nu_p| \approx \mathcal{O}\left(L\left[\frac{p-1}{N_x}\right]^2\right)$, let $K_1 \in \mathbb{R}$ denote the constant such that,

$$|1-\nu_p| \leq K_1 L\left[\frac{p-1}{N_x}\right]^2, \quad \text{for } p=2,\ldots,\frac{N_x+1}{2}. \tag{B.3}$$

Similarly, (4.40) determines that $\xi_p \approx \mathcal{O}(1)$ for $p=2,\ldots,\frac{N_x+1}{2}$, so let $K_2 \in \mathbb{R}$ be defined such that,

$$\xi_p \leq K_2, \quad \text{for } p=2,\ldots,\frac{N_x+1}{2}. \tag{B.4}$$

Along with the tables of numerical results, the analytical orders of convergence to zero for S_k, $k=1,\ldots,6$ for the Upwind scheme, are calculated using (B.3) and (B.4).

B.1 The orders of convergence for S_1...

B.1.1 with respect to N_x

Upwind Scheme: $\alpha_1 = \log\left(\frac{E_{1,3N_x}}{E_{1,N_x}}\right) / \log(3)$										
γ	$N_x = 3^\gamma$ \ r	0	1	2	3	4	5	6	7	$r \gg 1$
2	9	3.6315×10^{-4}	-1.8458	-2.7700	-2.9371	-2.9608	-2.9651	-2.9660	-2.9662	-2.9662
3	27	2.7423×10^{-6}	-1.9539	-2.9335	-2.9957	-2.9978	-2.9979	-2.9980	-2.9980	-2.9980
4	81	3.3971×10^{-8}	-1.9851	-2.9788	-2.9996	-2.9998	-2.9998	-2.9998	-2.9998	-2.9998
5	243	4.1954×10^{-10}	-1.9951	-2.9930	-3.0000	-3.0000	-3.0000	-3.0000	-3.0000	-3.0000
6	729	5.1800×10^{-12}	-1.9984	-2.9977	-3.0000	-3.0000	-3.0000	-3.0000	-3.0000	-3.0000

Table B.1: *The numerical orders of convergence to zero with respect to N_x for S_1, denoted by α_1, using the Upwind scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$ and $r \gg 1$, by considering $N_x = 3^\gamma$ for $\gamma = 2, \ldots, 7$ and fixed $L = 4$ and calculating them through α_1 as in Equation B.2. The values of N_x listed in the table are the smaller of the two values of N_x required to generate α_1. The other is identified by multiplying the listed value for N_x by three.*

Preissman Box Scheme: $\alpha_1 = \log\left(\frac{E_{1,3N_x}}{E_{1,N_x}}\right)/\log(3)$										
γ	$N_x = 3^\gamma$ \ r	0	1	2	3	4	5	6	7	$r \gg 1$
2	9	-2.4101×10^{-2}	-2.0093	-3.9363	-4.9749	-5.0571	-5.0522	-5.0502	-5.0497	-5.0496
3	27	-1.2880×10^{-3}	-2.0008	-3.9795	-4.9951	-5.0077	-5.0058	-5.0056	-5.0055	-5.0055
4	81	-1.4312×10^{-4}	-2.0001	-3.9932	-4.9987	-5.0009	-5.0007	-5.0006	-5.0006	-5.0006
5	243	-1.5900×10^{-5}	-2.0000	-3.9977	-4.9996	-5.0001	-5.0001	-5.0001	-5.0001	-5.0001
6	729	-1.7666×10^{-6}	-2.0000	-3.9992	-4.9999	-5.0000	-5.0000	-5.0000	-5.0000	-5.0000

Table B.2: *The numerical orders of convergence to zero with respect to N_x for S_1, denoted by α_1, using the Preissman Box scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$ and $r \gg 1$, by considering $N_x = 3^\gamma$ for $\gamma = 2, \ldots, 7$ and fixed $L = 4$ and calculating them through α_1 as in Equation B.2. The values of N_x listed in the table are the smaller of the two values of N_x required to generate α_1. The other is identified by multiplying the listed value for N_x by three.*

Lax-Wendroff Scheme: $\alpha_1 = \log\left(\frac{E_{1,3N_x}}{E_{1,N_x}}\right)/\log(3)$										
γ	$N_x = 3^\gamma$ \ r	0	1	2	3	4	5	6	7	$r \gg 1$
2	9	-2.4966×10^{-3}	-2.0017	-3.8678	-4.7934	-4.9364	-4.9530	-4.9556	-4.9562	-4.9563
3	27	-2.5276×10^{-4}	-2.0002	-3.9598	-4.9374	-4.9929	-4.9949	-4.9952	-4.9952	-4.9952
4	81	-2.7041×10^{-5}	-1.9914	-1.1592	1.2364	1.5877	1.6437	1.6556	-1.6584	1.6593
5	243	-4.3907×10^{-6}	-2.0058	-4.9212	-6.9921	-8.9490	-10.6354	-11.4681	-11.6350	-11.6587
6	729	-7.2463×10^{-7}	-2.0018	-4.7466	-6.8775	-7.4546	-6.0051	-5.1869	-5.0228	-5.0000
7	6561	-1.7562×10^{-7}	-2.0007	-4.5465	-6.5158	-5.1804	-5.0026	-5.0000	-5.0000	-5.0000
8	19683	3.7042×10^{-8}	-2.0002	-4.2080	-5.6001	-5.0027	-5.0000	-5.0000	-5.0000	-5.0000
9	59049	6.7067×10^{-9}	-2.0000	-3.9608	-5.1320	-5.0001	-5.0000	-5.0000	-5.0000	-5.0000
10	177147	-4.8417×10^{-9}	-2.0000	-4.0233	-5.0600	-5.0000	-5.0000	-5.0000	-5.0000	-5.0000
11	531441	-2.4129×10^{-9}	-2.0000	-4.0180	-5.0196	-5.0000	-5.0000	-5.0000	-5.0000	-5.0000

Table B.3: *The numerical orders of convergence to zero with respect to N_x for S_1, denoted by α_1, using the Lax-Wendroff scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$ and $r \gg 1$, by considering $N_x = 3^\gamma$ for $\gamma = 2, \ldots, 12$ and fixed $L = 4$ and calculating them through α_1 as in Equation B.2. The values of N_x listed in the table are the smaller of the two values of N_x required to generate α_1. The other is identified by multiplying the listed value for N_x by three.*

B.1.2 with respect to L

Upwind Scheme: $\beta_1 = \log\left(\frac{E_{1,2L}}{E_{1,L}}\right)/\log(2)$										
δ	$L = 2^\delta$ \ r	0	1	2	3	4	5	6	7	$r \gg 1$
0	1	1.6064	1.8588	1.9999	2.0000	2.0000	2.0000	2.0000	2.0000	2.0000
1	2	1.4069	1.7857	1.9998	2.0000	2.0000	2.0000	2.0000	2.0000	2.0000
2	4	1.1964	1.7077	1.9996	2.0000	2.0000	2.0000	2.0000	2.0000	2.0000
3	8	1.0085	1.6385	1.9994	2.0000	2.0000	2.0000	2.0000	2.0000	2.0000
4	16	8.6031×10^{-1}	1.5856	1.9990	2.0000	2.0000	2.0000	2.0000	2.0000	2.0000
5	32	7.5181×10^{-1}	1.5497	1.9985	2.0000	2.0000	2.0000	2.0000	2.0000	2.0000
6	64	6.7529×10^{-1}	1.5271	1.9979	2.0000	2.0000	2.0000	2.0000	2.0000	2.0000
7	128	6.2211×10^{-1}	1.5235	1.9970	2.0000	2.0000	2.0000	2.0000	2.0000	2.0000
8	256	5.8528×10^{-1}	1.5054	1.9957	2.0000	2.0000	2.0000	2.0000	2.0000	2.0000

Table B.4: *The numerical orders of convergence to zero with respect to L for S_1, denoted by β_1, using the Upwind scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$ and $r \gg 1$, by considering $L = 2^\delta$ for $\delta = 0, \ldots, 9$ and fixed $N_x = 3^7$ and calculating β_1 similarly to Equation B.2. The values of L listed in the table are the smaller of the two values of L required to generate β_1. The other is identified by multiplying the listed value for L by two.*

Preissman Box Scheme: $\beta_1 = \log\left(\frac{E_{1,2L}}{E_{1,L}}\right) / \log(2)$										
δ	r / $L = 2^\delta$	0	1	2	3	4	5	6	7	$r \gg 1$
0	1	1.6434	1.7301	1.8643	1.9999	2.0000	2.0000	2.0000	2.0000	2.0000
1	16	1.0341	1.3152	1.6898	1.9997	2.0000	2.0000	2.0000	2.0000	2.0000
2	4	7.0180×10^{-1}	1.0989	1.6211	1.9996	2.0000	2.0000	2.0000	2.0000	2.0000
3	8	5.9403×10^{-1}	1.0434	1.6227	1.9996	2.0000	2.0000	2.0000	2.0000	2.0000
4	16	5.1674×10^{-1}	1.0141	1.6319	1.9995	2.0000	2.0000	2.0000	2.0000	2.0000
5	32	4.6438×10^{-1}	1.0012	1.6417	1.9995	2.0000	2.0000	2.0000	2.0000	2.0000
6	64	4.2899×10^{-1}	9.9662×10^{-1}	1.6495	1.9994	2.0000	2.0000	2.0000	2.0000	2.0000
7	128	4.0457×10^{-1}	9.9573×10^{-1}	1.6551	1.9992	2.0000	2.0000	2.0000	2.0000	2.0000
8	256	3.8727×10^{-1}	9.9622×10^{-1}	1.6588	1.9991	2.0000	2.0000	2.0000	2.0000	2.0000

Table B.5: *The numerical orders of convergence to zero with respect to L for S_1, denoted by β_1, using the Preissman Box scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$ and $r \gg 1$, by considering $L = 2^\delta$ for $\delta = 0, \ldots, 9$ and fixed $N_x = 3^7$ and calculating β_1 similarly to Equation B.2. The values of L listed in the table are the smaller of the two values of L required to generate β_1. The other is identified by multiplying the listed value for L by two.*

Lax-Wendroff Scheme: $\beta_1 = \log\left(\frac{E_{1,2L}}{E_{1,L}}\right) / \log(2)$										
δ	r / $L = 2^\delta$	0	1	2	3	4	5	6	7	$r \gg 1$
0	1	1.5217	1.6882	3.6375×10^{-2}	-2.4638×10^{-2}	5.9916×10^{-1}	1.9370	1.9992	2.0000	2.0000
1	2	1.1287	1.4803	3.4093×10^{-2}	-2.3903×10^{-1}	1.2019	1.9831	1.9998	2.0000	2.0000
2	4	9.2646×10^{-1}	1.3559	7.9006×10^{-1}	3.4459×10^{-1}	1.7913	1.9969	2.0000	2.0000	2.0000
3	8	7.6918×10^{-1}	1.2501	1.2351	7.8452×10^{-1}	1.9392	1.9992	2.0000	2.0000	2.0000
4	16	6.4318×10^{-1}	1.1645	1.5309	1.4229	1.9844	1.9998	2.0000	2.0000	2.0000
5	32	5.5251×10^{-1}	1.1037	1.6409	1.8109	1.9961	2.0000	2.0000	2.0000	2.0000
6	64	4.8993×10^{-1}	1.0639	1.6702	1.9479	1.9990	2.0000	2.0000	2.0000	2.0000
7	128	4.4692×10^{-1}	1.0390	1.6747	1.9859	1.9998	2.0000	2.0000	2.0000	2.0000
8	256	4.1700×10^{-1}	1.0236	1.6731	1.9955	2.0000	2.0000	2.0000	2.0000	2.0000

Table B.6: *The numerical orders of convergence to zero with respect to L for S_1, denoted by β_1, using the Lax-Wendroff scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$ and $r \gg 1$, by considering $L = 2^\delta$ for $\delta = 0, \ldots, 9$ and fixed $N_x = 3^7$ and calculating β_1 similarly to Equation B.2. The values of L listed in the table are the smaller of the two values of L required to generate β_1. The other is identified by multiplying the listed value for L by two.*

B.1.3 analytically for the Upwind scheme

The analytical order of convergence of S_1, for the Upwind scheme, is calculated in the following.

$$
\begin{aligned}
E_1 &= |S_1 - 0| \\
&= \sum_{p=2}^{\frac{N_x+1}{2}} \frac{|1-\nu_p|^2}{(p-1)^{2(r+1)}} \\
&\leq \frac{K_1^2 L^2}{N_x^4} \sum_{p=1}^{\frac{N_x-1}{2}} p^{2-2r}, \\
&= \begin{cases} \frac{K_1^2 L^2}{6N_x^4}\left(\frac{N_x-1}{2}\right)\left[2\left(\frac{N_x-1}{2}\right)^2 + 3\left(\frac{N_x-1}{2}\right) + 1\right], & \text{for } r = 0, \\ \frac{K_1^2 L^2}{N_x^4}\left(\frac{N_x-1}{2}\right), & \text{for } r = 1, \\ \frac{K_1^2 L^2}{N_x^4} \sum_{p=1}^{\frac{N_x-1}{2}} \frac{1}{p^{2r-1}}, & \text{for } r \geq 2, \end{cases} \\
&\quad \text{where } \sum_{p=1}^{\frac{N_x-1}{2}} p^{2-2r} \text{ is calculated using [97] for } r = 0, 1, \\
&< \begin{cases} \frac{K_1^2 L^2}{24} N_x^{-1}, & \text{for } r = 0, \text{ as } N_x - 1 < N_x, \\ \frac{K_1^2 L^2}{2} N_x^{-3}, & \text{for } r = 1, \text{ as } N_x - 1 < N_x, \\ K_1^2 L^2 \zeta(2r-2) N_x^{-4}, & \text{for } r \geq 2, \end{cases}
\end{aligned}
$$

$$
\Rightarrow \quad 2D_2^2 N_x \sum_{p=2}^{\frac{N_x+1}{2}} \frac{|1-\nu_p|^2}{(p-1)^{2(r+1)}} < \begin{cases} \frac{D_2^2 K_1^2 L^2}{12} N_x^0, & \text{for } r = 0, \\ D_2^2 K_1^2 L^2 N_x^{-2}, & \text{for } r = 1, \\ 2D_2^2 K_1^2 L^2 \zeta(2r-2) N_x^{-3}, & \text{for } r \geq 2, \end{cases} \tag{B.5}
$$

$$
\Rightarrow \quad 2D_2^2 N_x \sum_{p=2}^{\frac{N_x+1}{2}} \frac{|1-\nu_p|^2}{(p-1)^{2(r+1)}} = \begin{cases} \mathcal{O}(L^2 N_x^{-2r}), & \text{for } r = 0, 1, \\ \mathcal{O}(L^2 N_x^{-3}), & \text{for } r \geq 2. \end{cases} \tag{B.6}
$$

B.2 The orders of convergence for S_2...

B.2.1 with respect to N_x

Upwind Scheme: $\alpha_2 = \log\left(\frac{E_{2,3N_x}}{E_{2,N_x}}\right) / \log(3)$									
γ	$N_x = 3^\gamma$ \ r	0	1	2	3	4	5	6	7
2	9	-7.8378×10^{-5}	-1.9996	-3.9764	-5.8458	-7.4594	-8.7700	-9.8924	-10.9371
3	27	-5.2357×10^{-6}	-2.0000	-3.9974	-5.9539	-7.6628	-8.9335	-9.9870	-10.9957
4	81	-6.4518×10^{-8}	-2.0000	-3.9997	-5.9851	-7.7545	-8.9788	-9.9984	-10.9996
5	243	-7.9636×10^{-10}	-2.0000	-4.0000	-5.9951	-7.8068	-8.9930	-9.9998	-11.0000
6	729	-9.8321×10^{-12}	-2.0000	-4.0000	-5.9984	-7.8407	-8.9977	-10.0000	-11.0000

Table B.7: *The numerical orders of convergence to zero with respect to N_x for S_2, denoted by α_2, using the Upwind scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$, by considering $N_x = 3^\gamma$ for $\gamma = 2, \ldots, 7$ and fixed $L = 4$ and calculating them through α_2 as in Equation B.2. The values of N_x listed in the table are the smaller of the two values of N_x required to generate α_2. The other is identified by multiplying the listed value for N_x by three.*

Preissman Box Scheme: $\alpha_2 = \log\left(\frac{E_{2,3N_x}}{E_{2,N_x}}\right)/\log(3)$									
γ	$N_x = 3^\gamma$ \ r	0	1	2	3	4	5	6	7
2	9	-3.3286×10^{-2}	-2.0241	-4.0160	-6.0093	-7.9982	-9.9363	-11.6515	-12.9749
3	27	-1.4632×10^{-3}	-2.0013	-4.0011	-6.0008	-7.9999	-9.9795	-11.7493	-12.9951
4	81	-1.5988×10^{-4}	-2.0001	-4.0001	-6.0001	-8.0000	-9.9932	-11.8037	-12.9987
5	243	-1.7731×10^{-5}	-2.0000	-4.0000	-6.0000	-8.0000	-9.9977	-11.8386	-12.9996
6	729	-1.9697×10^{-6}	-2.0000	-4.0000	-6.0000	-8.0000	-9.9993	-11.8630	-12.9999

Table B.8: *The numerical orders of convergence to zero with respect to N_x for S_2, denoted by α_2, using the Preissman Box scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$, by considering $N_x = 3^\gamma$ for $\gamma = 2, \ldots, 7$ and fixed $L = 4$ and calculating them through α_2 as in Equation B.2. The values of N_x listed in the table are the smaller of the two values of N_x required to generate α_2. The other is identified by multiplying the listed value for N_x by three.*

Lax-Wendroff Scheme: $\alpha_2 = \log\left(\frac{E_{2,3N_x}}{E_{2,N_x}}\right)/\log(3)$									
γ	$N_x = 3^\gamma$ \ r	0	1	2	3	4	5	6	7
2	9	-1.8236×10^{-3}	-2.0025	-4.0027	-6.0017	-7.9831	-9.8678	-11.4922	-12.7934
3	27	-2.4198×10^{-4}	-2.0003	-4.0002	-6.0002	-7.9981	-9.9598	-11.6751	-12.9374
4	81	-2.7439×10^{-5}	-2.0000	-3.9999	-5.9914	-7.6082	-7.1592	-6.5195	-6.7636
5	243	-3.0729×10^{-6}	-2.0000	-4.0001	-6.0058	-8.2415	-10.9213	-12.9922	-14.9921
6	729	-3.4445×10^{-7}	-2.0000	-4.0000	-6.0018	-8.0925	-10.7466	-12.9106	-14.8775
7	6561	-3.9481×10^{-8}	-2.0000	-4.0000	-6.0007	-8.0355	-10.5465	-12.8393	-14.5158
8	19683	-4.6036×10^{-9}	-2.0000	-4.0000	-6.0002	-8.0090	-10.2080	-12.4170	-13.6001
9	59049	-3.6991×10^{-10}	-2.0000	-4.0000	-6.0000	-7.9984	-9.9608	-11.8959	-13.1320
10	177147	-1.1795×10^{-10}	-2.0000	-4.0000	-6.0000	-8.0010	-10.0233	-12.0281	-13.0600
11	531441	-3.5133×10^{-11}	-2.0000	-4.0000	-6.0000	-8.0000	-10.0180	-12.0201	-13.0196

Table B.9: *The numerical orders of convergence to zero with respect to N_x for S_2, denoted by α_2, using the Upwind scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$, by considering $N_x = 3^\gamma$ for $\gamma = 2, \ldots, 12$ and fixed $L = 4$ and calculating them through α_2 as in Equation B.2. The values of N_x listed in the table are the smaller of the two values of N_x required to generate α_2. The other is identified by multiplying the listed value for N_x by three.*

B.2.2 with respect to L

Upwind Scheme: $\beta_2 = \log\left(\frac{E_{2,2L}}{E_{2,L}}\right)/\log(2)$									
δ	$L = 2^\delta$ \ r	0	1	2	3	4	5	6	7
0	1	1.4971	1.6064	1.7270	1.8588	1.9797	1.9999	2.0000	2.0000
1	2	1.2456	1.4069	1.5870	1.7857	1.9684	1.9998	2.0000	2.0000
2	4	9.8443×10^{-1}	1.1964	1.4383	1.7077	1.9556	1.9996	2.0000	2.0000
3	8	7.5586×10^{-1}	1.0085	1.3051	1.6385	1.9431	1.9994	2.0000	2.0000
4	16	5.7888×10^{-1}	8.6031×10^{-1}	1.2011	1.5856	1.9322	1.9990	2.0000	2.0000
5	32	4.5067×10^{-1}	7.5181×10^{-1}	1.1272	1.5497	1.9230	1.9985	2.0000	2.0000
6	64	3.5982×10^{-1}	6.7529×10^{-1}	1.0780	1.5271	1.9149	1.9979	2.0000	2.0000
7	128	2.9509×10^{-1}	6.2211×10^{-1}	1.0467	1.5135	1.9074	1.9970	2.0000	2.0000
8	256	2.4804×10^{-1}	5.8528×10^{-1}	1.0274	1.5054	1.8998	1.9957	1.9999	2.0000

Table B.10: *The numerical orders of convergence to zero with respect to L for S_2, denoted by β_2, using the Upwind scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$, by considering $L = 2^\delta$ for $\delta = 0, \ldots, 9$ and fixed $N_x = 3^7$ and calculating β_2 similarly to Equation B.2. The values of L listed in the table are the smaller of the two values of L required to generate β_2. The other is identified by multiplying the listed value for L by two.*

Preissman Box Scheme: $\beta_2 = \log\left(\frac{E_{2,2L}}{E_{2,L}}\right) / \log(2)$									
δ	$L = 2^\delta$ \ r	0	1	2	3	4	5	6	7
0	1	1.6091	1.6434	1.6831	1.7301	1.7877	1.8643	1.9730	1.9999
1	16	9.1363×10^{-1}	1.0341	1.1669	1.3152	1.4845	1.6898	1.9434	1.9997
2	4	5.4107×10^{-1}	7.0180×10^{-1}	8.8742×10^{-1}	1.0989	1.3396	1.6211	1.9356	1.9996
3	8	4.2098×10^{-1}	5.9403×10^{-1}	8.0241×10^{-1}	1.0434	1.3153	1.6227	1.9388	1.9996
4	16	3.3180×10^{-1}	5.1674×10^{-1}	7.4714×10^{-1}	1.0141	1.3096	1.6319	1.9416	1.9995
5	32	2.6925×10^{-1}	4.6438×10^{-1}	7.1392×10^{-1}	1.0012	1.3124	1.6417	1.9433	1.9995
6	64	2.2532×10^{-1}	4.2899×10^{-1}	6.9449×10^{-1}	9.9662×10^{-1}	1.3173	1.6495	1.9439	1.9994
7	128	1.9362×10^{-1}	4.0457×10^{-1}	6.8318×10^{-1}	9.9573×10^{-1}	1.3220	1.6551	1.9436	1.9992
8	256	1.6992×10^{-1}	3.8727×10^{-1}	6.7655×10^{-1}	9.9622×10^{-1}	1.3256	1.6588	1.9426	1.9991

Table B.11: *The numerical orders of convergence to zero with respect to L for S_2, denoted by β_2, using the Preissman Box scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$, by considering $L = 2^\delta$ for $\delta = 0, \ldots, 9$ and fixed $N_x = 3^7$ and calculating β_2 similarly to Equation B.2. The values of L listed in the table are the smaller of the two values of L required to generate β_2. The other is identified by multiplying the listed value for L by two.*

Lax-Wendroff Scheme: $\beta_2 = \log\left(\frac{E_{2,2L}}{E_{2,L}}\right) / \log(2)$									
δ	$L = 2^\delta$ \ r	0	1	2	3	4	5	6	7
0	1	1.4455	1.5217	1.6072	1.6882	1.1797	3.6375×10^{-2}	-3.8606×10^{-2}	-2.4638×10^{-2}
1	2	9.6510×10^{-1}	1.1287	1.3029	1.4803	1.4277	3.4093×10^{-2}	-2.8286×10^{-1}	-2.3903×10^{-1}
2	4	7.3301×10^{-1}	9.2646×10^{-1}	1.1363	1.3559	1.5227	7.9006×10^{-1}	2.0183×10^{-1}	3.4459×10^{-1}
3	8	5.6114×10^{-1}	7.6918×10^{-1}	1.0018	1.2501	1.4858	1.2351	4.5284×10^{-1}	7.8452×10^{-1}
4	16	4.2722×10^{-1}	6.4318×10^{-1}	8.9319×10^{-1}	1.1645	1.4392	1.5309	1.0377	1.4229
5	32	3.3318×10^{-1}	5.5251×10^{-1}	8.1534×10^{-1}	1.1037	1.4014	1.6409	1.5798	1.8109
6	64	2.6883×10^{-1}	4.8993×10^{-1}	7.6292×10^{-1}	1.0639	1.3756	1.6702	1.8358	1.9479
7	128	2.2409×10^{-1}	4.4692×10^{-1}	7.2864×10^{-1}	1.0390	1.3592	1.6747	1.9160	1.9859
8	256	1.9195×10^{-1}	4.1700×10^{-1}	7.0649×10^{-1}	1.0236	1.3492	1.6731	1.9359	1.9955

Table B.12: *The numerical orders of convergence to zero with respect to L for S_2, denoted by β_2, using the Lax-Wendroff scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$, by considering $L = 2^\delta$ for $\delta = 0, \ldots, 9$ and fixed $N_x = 3^7$ and calculating β_2 similarly to Equation B.2. The values of L listed in the table are the smaller of the two values of L required to generate β_2. The other is identified by multiplying the listed value for L by two.*

B.2.3 analytically for the Upwind scheme

The analytical order of convergence of S_2, for the Upwind scheme, is calculated in the following.

$$
\begin{aligned}
E_2 &= |S_2 - 0| \\
&= \sum_{p=2}^{\frac{N_x+1}{2}} \frac{|1-\nu_p|^2}{(p-1)^{r+1}} \\
&\leq \frac{K_1^2 L^2}{N_x^4} \sum_{p=1}^{\frac{N_x-1}{2}} p^{3-r}, \\
&< \begin{cases}
\frac{3K_1^2}{8} L^2 N_x^0, & \text{for } r = 0, \\
\frac{K_1^2}{24} L^2 N_x^{-1}, & \text{for } r = 1, \text{ as } N_x - 1 < N_x, \\
\frac{3K_1^2}{4} L^2 N_x^{-2}, & \text{for } r = 2, \text{ as } N_x - 1 < N_x, \\
\frac{K_1^2}{2} L^2 N_x^{-3}, & \text{for } r = 3, \text{ as } N_x - 1 < N_x, \\
K_1^2 L^2 N_x^{-4} \left(\frac{1}{N_x - 1} + \log(N_x - 1) - \log(2) + \gamma \right), & \text{for } r = 4, \\
K_1^2 \zeta(r-3) L^2 N_x^{-4}, & \text{for } r \geq 5,
\end{cases}
\end{aligned}
\tag{B.7}
$$

where $\sum_{p=1}^{\frac{N_x-1}{2}} p^{3-r}$ is calculated using [97] for $r \leq 3$. Here we have used the Riemann zeta function to bound $\sum_{p=1}^{\frac{N_x-1}{2}} p^{3-r}$ for $r \geq 5$, as the series is convergent. However, when $r = 4$, the series is divergent. In this instance we have used the result, [98],

$$
\sum_{p=1}^{n} \frac{1}{p} = H_n < \frac{1}{2n} + \log(n) + \gamma, \tag{B.8}
$$

where H_n is the nth harmonic number for $n \in \mathbb{N}$ and γ is the Euler-Mascheroni constant [98]. This give that,

$$
\begin{aligned}
\sum_{p=1}^{\frac{N_x-1}{2}} \frac{1}{p} &< \frac{1}{N_x - 1} + \log(N_x - 1) - \log(2) + \gamma, \\
&< (2 + \gamma) \log(N_x). \qquad \text{(B.9)}
\end{aligned}
\tag{B.10}
$$

Therefore,

$$\frac{2D_2D_3}{N_x^r}\sum_{p=2}^{\frac{N_x+1}{2}}\frac{|1-\nu_p|^2}{(p-1)^{r+1}} < \begin{cases} \frac{3D_2D_3K_1^2}{4}L^2N_x^0, & \text{for } r=0, \\ \frac{D_2D_3K_1^2}{12}L^2N_x^{-2}, & \text{for } r=1, \\ \frac{3D_2D_3K_1^2}{2}L^2N_x^{-4}, & \text{for } r=2, \\ D_2D_3K_1^2L^2N_x^{-6}, & \text{for } r=3, \\ 2D_2D_3K_1^2(2+\gamma)L^2N_x^{-8}\log(N_x), & \text{for } r=4, \\ 2D_2D_3K_1^2\zeta(r-3)L^2N_x^{-4-r}, & \text{for } r\geq 5. \end{cases} \tag{B.11}$$

$$\Rightarrow \quad \frac{2D_2D_3}{N_x^r}\sum_{p=2}^{\frac{N_x+1}{2}}\frac{|1-\nu_p|^2}{(p-1)^{r+1}} = \begin{cases} \mathcal{O}(L^2N_x^{-2r}), & \text{for } r=0,\ldots,3, \\ \mathcal{O}(L^2N_x^{-2r}\log(N_x)), & \text{for } r=4, \\ \mathcal{O}(L^2N_x^{-4-r}), & \text{for } r\geq 5. \end{cases} \tag{B.12}$$

B.3 The orders of convergence for S_3...

B.3.1 with respect to N_x

Upwind Scheme: $\alpha_3 = \log\left(\frac{E_{3,3N_x}}{E_{3,N_x}}\right) / \log(3)$									
γ	$N_x = 3^\gamma$ \ r	0	1	2	3	4	5	6	7
2	9	4.4162×10^{-4}	−1.9996	−3.9996	−5.9996	−7.9996	−9.9996	−11.9996	−13.9996
3	27	-1.8189×10^{-9}	−2.0000	−4.0000	−6.0000	−8.0000	−10.0000	−12.0000	−14.0000
4	81	-2.7164×10^{-13}	−2.0000	−4.0000	−6.0000	−8.0000	−10.0000	−12.0000	−14.0000
5	243	6.0634×10^{-16}	−2.0000	−4.0000	−6.0000	−8.0000	−10.0000	−12.0000	−14.0000
6	729	4.0423×10^{-16}	−2.0000	−4.0000	−6.0000	−8.0000	−10.0000	−12.0000	−14.0000

Table B.13: *The numerical orders of convergence to zero with respect to N_x for S_3, denoted by α_3, using the Upwind scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$, by considering $N_x = 3^\gamma$ for $\gamma = 2, \ldots, 7$ and fixed $L = 4$ and calculating them through α_3 as in Equation B.2. The values of N_x listed in the table are the smaller of the two values of N_x required to generate α_3. The other is identified by multiplying the listed value for N_x by three.*

Preissman Box Scheme: $\alpha_3 = \log\left(\frac{E_{3,3N_x}}{E_{3,N_x}}\right)/\log(3)$									
γ	$N_x = 3^\gamma$ \ r	0	1	2	3	4	5	6	7
2	9	-4.2781×10^{-2}	−2.0428	−4.0428	−6.0428	−8.0428	−10.0428	−12.0428	−14.0428
3	27	-1.5692×10^{-3}	−2.0016	−4.0016	−6.0016	−8.0016	−10.0016	−12.0016	−14.0016
4	81	-1.6772×10^{-4}	−2.0002	−4.0002	−6.0002	−8.0002	−10.0002	−12.0002	−14.0002
5	243	-1.8558×10^{-5}	−2.0000	−4.0000	−6.0000	−8.0000	−10.0000	−12.0000	−14.0000
6	729	-2.0610×10^{-6}	−2.0000	−4.0000	−6.0000	−8.0000	−10.0000	−12.0000	−14.0000

Table B.14: *The numerical orders of convergence to zero with respect to N_x for S_3, denoted by α_3, using the Preissman Box scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$, by considering $N_x = 3^\gamma$ for $\gamma = 2, \ldots, 7$ and fixed $L = 4$ and calculating them through α_3 as in Equation B.2. The values of N_x listed in the table are the smaller of the two values of N_x required to generate α_3. The other is identified by multiplying the listed value for N_x by three.*

Lax-Wendroff Scheme: $\alpha_3 = \log\left(\frac{E_{3,3N_x}}{E_{3,N_x}}\right)/\log(3)$									
γ	r / $N_x = 3^\gamma$	0	1	2	3	4	5	6	7
2	9	-2.3042×10^{-4}	-2.0002	-4.0002	-6.0002	-8.0002	-10.0002	-12.0002	-14.0002
3	27	-1.4104×10^{-4}	-2.0001	-4.0001	-6.0001	-8.0001	-10.0001	-12.0001	-14.0001
4	81	-1.5771×10^{-5}	-2.0000	-4.0000	-6.0000	-8.0000	-10.0000	-12.0000	-14.0000
5	243	-1.7536×10^{-6}	-2.0000	-4.0000	-6.0000	-8.0000	-10.0000	-12.0000	-14.0000
6	729	-1.9489×10^{-6}	-2.0000	-4.0000	-6.0000	-8.0000	-10.0000	-12.0000	-14.0000

Table B.15: *The numerical orders of convergence to zero with respect to N_x for S_3, denoted by α_3, using the Lax-Wendroff scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$, by considering $N_x = 3^\gamma$ for $\gamma = 2, \ldots, 7$ and fixed $L = 4$ and calculating them through α_3 as in Equation B.2. The values of N_x listed in the table are the smaller of the two values of N_x required to generate α_3. The other is identified by multiplying the listed value for N_x by three.*

B.3.2 with respect to L

Upwind Scheme: $\beta_3 = \log\left(\frac{E_{3,2L}}{E_{3,L}}\right)/\log(2)$									
δ	$L = 2^\delta$ \ r	0	1	2	3	4	5	6	7
0	1	1.3980	1.3980	1.3980	1.3980	1.3980	1.3980	1.3980	1.3980
1	2	1.1024	1.1024	1.1024	1.1024	1.1024	1.1024	1.1024	1.1024
2	4	8.0255×10^{-1}	8.0255×10^{-1}	8.0255×10^{-1}	8.0255×10^{-1}	8.0255×10^{-1}	8.0255×10^{-1}	8.0255×10^{-1}	8.0255×10^{-1}
3	8	5.5018×10^{-1}	5.5018×10^{-1}	5.5018×10^{-1}	5.5018×10^{-1}	5.5018×10^{-1}	5.5018×10^{-1}	5.5018×10^{-1}	5.5018×10^{-1}
4	16	3.6530×10^{-1}	3.6530×10^{-1}	3.6530×10^{-1}	3.6530×10^{-1}	3.6530×10^{-1}	3.6530×10^{-1}	3.6530×10^{-1}	3.6530×10^{-1}
5	32	2.4051×10^{-1}	2.4051×10^{-1}	2.4051×10^{-1}	2.4051×10^{-1}	2.4051×10^{-1}	2.4051×10^{-1}	2.4051×10^{-1}	2.4051×10^{-1}
6	64	1.5921×10^{-1}	1.5921×10^{-1}	1.5921×10^{-1}	1.5921×10^{-1}	1.5921×10^{-1}	1.5921×10^{-1}	1.5921×10^{-1}	1.5921×10^{-1}
7	128	1.0656×10^{-1}	1.0656×10^{-1}	1.0656×10^{-1}	1.0656×10^{-1}	1.0656×10^{-1}	1.0656×10^{-1}	1.0656×10^{-1}	1.0656×10^{-1}
8	256	7.2171×10^{-2}	7.2171×10^{-2}	7.2171×10^{-2}	7.2171×10^{-2}	7.2171×10^{-2}	7.2171×10^{-2}	7.2171×10^{-2}	7.2171×10^{-2}

Table B.16: *The numerical orders of convergence to zero with respect to L for S_3, denoted by β_3, using the Upwind scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$, by considering $L = 2^\delta$ for $\delta = 0, \ldots, 9$ and fixed $N_x = 3^7$ and calculating β_3 similarly to Equation B.2. The values of L listed in the table are the smaller of the two values of L required to generate β_3. The other is identified by multiplying the listed value for L by two.*

Preissman Box Scheme: $\beta_3 = \log\left(\frac{E_{3,2L}}{E_{3,L}}\right) / \log(2)$									
δ	$L = 2^\delta$ \ r	0	1	2	3	4	5	6	7
0	1	1.5790	1.5790	1.5790	1.5790	1.5790	1.5790	1.5790	1.5790
1	16	8.0381×10^{-1}	8.0381×10^{-1}	8.0381×10^{-1}	8.0381×10^{-1}	8.0381×10^{-1}	8.0381×10^{-1}	8.0381×10^{-1}	8.0381×10^{-1}
2	4	4.0463×10^{-1}	4.0463×10^{-1}	4.0463×10^{-1}	4.0463×10^{-1}	4.0463×10^{-1}	4.0463×10^{-1}	4.0463×10^{-1}	4.0463×10^{-1}
3	8	2.8453×10^{-1}	2.8453×10^{-1}	2.8453×10^{-1}	2.8453×10^{-1}	2.8453×10^{-1}	2.8453×10^{-1}	2.8453×10^{-1}	2.8453×10^{-1}
4	16	1.9683×10^{-1}	1.9683×10^{-1}	1.9683×10^{-1}	1.9683×10^{-1}	1.9683×10^{-1}	1.9683×10^{-1}	1.9683×10^{-1}	1.9683×10^{-1}
5	32	1.3754×10^{-1}	1.3754×10^{-1}	1.3754×10^{-1}	1.3754×10^{-1}	1.3754×10^{-1}	1.3754×10^{-1}	1.3754×10^{-1}	1.3754×10^{-1}
6	64	9.8138×10^{-2}	9.8138×10^{-2}	9.8138×10^{-2}	9.8138×10^{-2}	9.8138×10^{-2}	9.8138×10^{-2}	9.8138×10^{-2}	9.8138×10^{-2}
7	128	7.1620×10^{-2}	7.1620×10^{-2}	7.1620×10^{-2}	7.1620×10^{-2}	7.1620×10^{-2}	7.1620×10^{-2}	7.1620×10^{-2}	7.1620×10^{-2}
8	256	5.3321×10^{-2}	5.3321×10^{-2}	5.3321×10^{-2}	5.3321×10^{-2}	5.3321×10^{-2}	5.3321×10^{-2}	5.3321×10^{-2}	5.3321×10^{-2}

Table B.17: *The numerical orders of convergence to zero with respect to L for S_3, denoted by β_3, using the Preissman Box scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$, by considering $L = 2^\delta$ for $\delta = 0, \ldots, 9$ and fixed $N_x = 3^7$ and calculating β_3 similarly to Equation B.2. The values of L listed in the table are the smaller of the two values of L required to generate β_3. The other is identified by multiplying the listed value for L by two.*

Lax-Wendroff Scheme: $\beta_3 = \log\left(\frac{E_{3,2L}}{E_{3,L}}\right) / \log(2)$									
δ	$L = 2^\delta$ \ r	0	1	2	3	4	5	6	7
0	1	1.3783	1.3783	1.3783	1.3783	1.3783	1.3783	1.3783	1.3783
1	2	8.1398×10^{-1}	8.1398×10^{-1}	8.1398×10^{-1}	8.1398×10^{-1}	8.1398×10^{-1}	8.1398×10^{-1}	8.1398×10^{-1}	8.1398×10^{-1}
2	4	5.6042×10^{-1}	5.6042×10^{-1}	5.6042×10^{-1}	5.6042×10^{-1}	5.6042×10^{-1}	5.6042×10^{-1}	5.6042×10^{-1}	5.6042×10^{-1}
3	8	3.8514×10^{-1}	3.8514×10^{-1}	3.8514×10^{-1}	3.8514×10^{-1}	3.8514×10^{-1}	3.8514×10^{-1}	3.8514×10^{-1}	3.8514×10^{-1}
4	16	2.5655×10^{-1}	2.5655×10^{-1}	2.5655×10^{-1}	2.5655×10^{-1}	2.5655×10^{-1}	2.5655×10^{-1}	2.5655×10^{-1}	2.5655×10^{-1}
5	32	1.7279×10^{-1}	1.7279×10^{-1}	1.7279×10^{-1}	1.7279×10^{-1}	1.7279×10^{-1}	1.7279×10^{-1}	1.7279×10^{-1}	1.7279×10^{-1}
6	64	1.1976×10^{-1}	1.1976×10^{-1}	1.1976×10^{-1}	1.1976×10^{-1}	1.1976×10^{-1}	1.1976×10^{-1}	1.1976×10^{-1}	1.1976×10^{-1}
7	128	8.5553×10^{-2}	8.5553×10^{-2}	8.5553×10^{-2}	8.5553×10^{-2}	8.5553×10^{-2}	8.5553×10^{-2}	8.5553×10^{-2}	8.5553×10^{-2}
8	256	6.2698×10^{-2}	6.2698×10^{-2}	6.2698×10^{-2}	6.2698×10^{-2}	6.2698×10^{-2}	6.2698×10^{-2}	6.2698×10^{-2}	6.2698×10^{-2}

Table B.18: *The numerical orders of convergence to zero with respect to L for S_3, denoted by β_3, using the Lax-Wendroff scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$, by considering $L = 2^\delta$ for $\delta = 0, \ldots, 9$ and fixed $N_x = 3^7$ and calculating β_3 similarly to Equation B.2. The values of L listed in the table are the smaller of the two values of L required to generate β_3. The other is identified by multiplying the listed value for L by two.*

B.3.3 analytically for the Upwind scheme

The analytical order of convergence of S_3, for the Upwind scheme, is calculated in the following.

$$
\begin{aligned}
E_3 &= |S_3 - 0| \\
&= \sum_{p=2}^{\frac{N_x+1}{2}} |1-\nu_p|^2 \\
&\leq \frac{K_1^2 L^2}{N_x^4} \sum_{p=1}^{\frac{N_x-1}{2}} p^4, \\
&= \frac{K_1^2 L^2}{30 N_x^4} \left(\frac{N_x-1}{2}\right) \left[6\left(\frac{N_x-1}{2}\right)^4 + 15\left(\frac{N_x-1}{2}\right)^3 + 10\left(\frac{N_x-1}{2}\right)^2 - 1\right], \\
&\quad \text{where } \sum_{p=1}^{\frac{N_x-1}{2}} p^4 \text{ is calculated using [97]}, \\
&< \frac{K_1^2 L^2}{160} N_x, \text{ as } N_x - 1 < N_x,
\end{aligned} \tag{B.13}
$$

$$
\Rightarrow \frac{2D_3^2}{N_x^{2r+1}} \sum_{p=2}^{\frac{N_x+1}{2}} |1-\nu_p|^2 < \frac{D_3^2 K_1^2}{80} L^2 N_x^{-2r}, \tag{B.14}
$$

$$
\Rightarrow \frac{2D_3^2}{N_x^{2r+1}} \sum_{p=2}^{\frac{N_x+1}{2}} |1-\nu_p|^2 = \mathcal{O}(L^2 N_x^{-2r}). \tag{B.15}
$$

B.4 The orders of convergence for S_4...

B.4.1 with respect to N_x

Upwind Scheme: $\alpha_4 = \log\left(\frac{E_{4,3N_x}}{E_{4,N_x}}\right) / \log(3)$								
γ	$N_x = 3^\gamma$ \ r	0	1	2	3	4	5	6	7
2	9	7.6321×10^{-5}	-1.9999	-3.9999	-5.9999	-8.0000	-9.9999	-12.0000	-14.0000
3	27	1.4148×10^{-15}	-2.0000	-4.0000	-6.0000	-8.0000	-10.0000	-12.0000	-14.0000
4	81	-7.0740×10^{-16}	-2.0000	-4.0000	-6.0000	-8.0000	-10.0000	-12.0000	-14.0000
5	243	4.2444×10^{-15}	-2.0000	-4.0000	-6.0000	-8.0000	-10.0000	-12.0000	-14.0000
6	729	-6.3666×10^{-15}	-2.0000	-4.0000	-6.0000	-8.0000	-10.0000	-12.0000	-14.0000

Table B.19: *The numerical orders of convergence to zero with respect to N_x for S_4, denoted by α_4, using the Upwind scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$, by considering $N_x = 3^\gamma$ for $\gamma = 2, \ldots, 7$ and fixed $L = 4$ and calculating them through α_4 as in Equation B.2. The values of N_x listed in the table are the smaller of the two values of N_x required to generate α_4. The other is identified by multiplying the listed value for N_x by three.*

Preissman Box Scheme: $\alpha_4 = \log\left(\frac{E_{4,3N_x}}{E_{4,N_x}}\right) / \log(3)$									
γ	$N_x = 3^\gamma$ \ r	0	1	2	3	4	5	6	7
2	9	-6.0200×10^{-4}	−2.0006	−4.0006	−6.0006	−8.0006	−10.0006	−12.0006	−14.0006
3	27	-4.7527×10^{-12}	−2.0000	−4.0000	−6.0000	−8.0000	−10.0000	−12.0000	−14.0000
4	81	-2.0211×10^{-16}	−2.0000	−4.0000	−6.0000	−8.0000	−10.0000	−12.0000	−14.0000
5	243	1.0106×10^{-15}	−2.0000	−4.0000	−6.0000	−8.0000	−10.0000	−12.0000	−14.0000
6	729	-2.0211×10^{-16}	−2.0000	−4.0000	−6.0000	−8.0000	−10.0000	−12.0000	−14.0000

Table B.20: *The numerical orders of convergence to zero with respect to N_x for S_4, denoted by α_4, using the Preissman Box scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$, by considering $N_x = 3^\gamma$ for $\gamma = 2, \ldots, 7$ and fixed $L = 4$ and calculating them through α_4 as in Equation B.2. The values of N_x listed in the table are the smaller of the two values of N_x required to generate α_4. The other is identified by multiplying the listed value for N_x by three.*

Lax-Wendroff Scheme: $\alpha_4 = \log\left(\frac{E_{4,3N_x}}{E_{4,N_x}}\right)/\log(3)$									
γ	r / $N_x = 3^\gamma$	0	1	2	3	4	5	6	7
2	9	1.7945×10^{-4}	-1.9998	-3.9998	-5.9998	-7.9998	-9.9998	-11.9998	-13.9998
3	27	-4.9781×10^{-13}	-2.0000	-4.0000	-6.0000	-8.0000	-10.0000	-12.0000	-14.0000
4	81	-8.3270×10^{-12}	-2.0000	-4.0000	-6.0000	-8.0000	-10.0000	-12.0000	-14.0000
5	243	-1.6289×10^{-11}	-2.0000	-4.0000	-6.0000	-8.0000	-10.0000	-12.0000	-14.0000
6	729	-1.9397×10^{-11}	-2.0000	-4.0000	-6.0000	-8.0000	-10.0000	-12.0000	-14.0000

Table B.21: *The numerical orders of convergence to zero with respect to N_x for S_4, denoted by α_4, using the Lax-Wendroff scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$, by considering $N_x = 3^\gamma$ for $\gamma = 2, \ldots, 7$ and fixed $L = 4$ and calculating them through α_4 as in Equation B.2. The values of N_x listed in the table are the smaller of the two values of N_x required to generate α_4. The other is identified by multiplying the listed value for N_x by three.*

MNIMC Scheme: $\alpha_4 = \log\left(\frac{E_{4,3N_x}}{E_{4,N_x}}\right)/\log(3)$									
γ	$N_x = 3^\gamma$ \ r	0	1	2	3	4	5	6	7
2	9	4.0423×10^{-16}	-2.0000	-4.0000	-6.0000	-8.0000	-10.0000	-12.0000	-14.0000
3	27	2.0211×10^{-16}	-2.0000	-4.0000	-6.0000	-8.0000	-10.0000	-12.0000	-14.0000
4	81	-2.0211×10^{-16}	-2.0000	-4.0000	-6.0000	-8.0000	-10.0000	-12.0000	-14.0000
5	243	-8.7919×10^{-15}	-2.0000	-4.0000	-6.0000	-8.0000	-10.0000	-12.0000	-14.0000
6	729	-7.2761×10^{-15}	-2.0000	-4.0000	-6.0000	-8.0000	-10.0000	-12.0000	-14.0000

Table B.22: *The numerical orders of convergence to zero with respect to N_x for S_4, denoted by α_4, using the MNIMC scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$, by considering $N_x = 3^\gamma$ for $\gamma = 2, \ldots, 7$ and fixed $L = 4$ and calculating them through α_4 as in Equation B.2. The values of N_x listed in the table are the smaller of the two values of N_x required to generate α_4. The other is identified by multiplying the listed value for N_x by three.*

B.4.2 with respect to L

Upwind Scheme: $\beta_4 = \log\left(\frac{E_{4,2L}}{E_{4,L}}\right)/\log(2)$									
δ	$L = 2^\delta$ \ r	0	1	2	3	4	5	6	7
0	1	-6.3871×10^{-1}	-6.3871×10^{-1}	-6.3871×10^{-1}	-6.3871×10^{-1}	-6.3871×10^{-1}	-6.3871×10^{-1}	-6.3871×10^{-1}	-6.3871×10^{-1}
1	2	6.8504×10^{-1}	6.8504×10^{-1}	6.8504×10^{-1}	6.8504×10^{-1}	6.8504×10^{-1}	6.8504×10^{-1}	6.8504×10^{-1}	6.8504×10^{-1}
2	4	4.7154×10^{-1}	4.7154×10^{-1}	4.7154×10^{-1}	4.7154×10^{-1}	4.7154×10^{-1}	4.7154×10^{-1}	4.7154×10^{-1}	4.7154×10^{-1}
3	8	3.1771×10^{-1}	3.1771×10^{-1}	3.1771×10^{-1}	3.1771×10^{-1}	3.1771×10^{-1}	3.1771×10^{-1}	3.1771×10^{-1}	3.1771×10^{-1}
4	16	2.1236×10^{-1}	2.1236×10^{-1}	2.1236×10^{-1}	2.1236×10^{-1}	2.1236×10^{-1}	2.1236×10^{-1}	2.1236×10^{-1}	2.1236×10^{-1}
5	32	1.4222×10^{-1}	1.4222×10^{-1}	1.4222×10^{-1}	1.4222×10^{-1}	1.4222×10^{-1}	1.4222×10^{-1}	1.4222×10^{-1}	1.4222×10^{-1}
6	64	9.5961×10^{-2}	9.5961×10^{-2}	9.5961×10^{-2}	9.5961×10^{-2}	9.5961×10^{-2}	9.5961×10^{-2}	9.5961×10^{-2}	9.5961×10^{-2}
7	128	6.5345×10^{-2}	6.5345×10^{-2}	6.5345×10^{-2}	6.5345×10^{-2}	6.5345×10^{-2}	6.5345×10^{-2}	6.5345×10^{-2}	6.5345×10^{-2}
8	256	4.4885×10^{-2}	4.4885×10^{-2}	4.4885×10^{-2}	4.4885×10^{-2}	4.4885×10^{-2}	4.4885×10^{-2}	4.4885×10^{-2}	4.4885×10^{-2}

Table B.23: *The numerical orders of convergence to zero with respect to L for S_4, denoted by β_4, using the Upwind scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$, by considering $L = 2^\delta$ for $\delta = 0, \ldots, 9$ and fixed $N_x = 3^7$ and calculating β_4 similarly to Equation B.2. The values of L listed in the table are the smaller of the two values of L required to generate β_4. The other is identified by multiplying the listed value for L by two.*

Preissman Box Scheme: $\beta_4 = \log\left(\frac{E_{4,2L}}{E_{4,L}}\right) / \log(2)$									
δ	$L = 2^\delta$ \ r	0	1	2	3	4	5	6	7
0	1	-1.1699	-1.1699	-1.1699	-1.1699	-1.1699	-1.1699	-1.1699	-1.1699
1	16	-1.9745×10^{-1}	-1.9745×10^{-1}	-1.9745×10^{-1}	-1.9745×10^{-1}	-1.9745×10^{-1}	-1.9745×10^{-1}	-1.9745×10^{-1}	-1.9745×10^{-1}
2	4	2.4088×10^{-3}	2.4088×10^{-3}	2.4088×10^{-3}	2.4088×10^{-3}	2.4088×10^{-3}	2.4088×10^{-3}	2.4088×10^{-3}	2.4088×10^{-3}
3	8	-1.3396×10^{-1}	-1.3396×10^{-1}	-1.3396×10^{-1}	-1.3396×10^{-1}	-1.3396×10^{-1}	-1.3396×10^{-1}	-1.3396×10^{-1}	-1.3396×10^{-1}
4	16	-2.1908×10^{-1}	-2.1908×10^{-1}	-2.1908×10^{-1}	-2.1908×10^{-1}	-2.1908×10^{-1}	-2.1908×10^{-1}	-2.1908×10^{-1}	-2.1908×10^{-1}
5	32	-2.6889×10^{-1}	-2.6889×10^{-1}	-2.6889×10^{-1}	-2.6889×10^{-1}	-2.6889×10^{-1}	-2.6889×10^{-1}	-2.6889×10^{-1}	-2.6889×10^{-1}
6	64	-2.9712×10^{-1}	-2.9712×10^{-1}	-2.9712×10^{-1}	-2.9712×10^{-1}	-2.9712×10^{-1}	-2.9712×10^{-1}	-2.9712×10^{-1}	-2.9712×10^{-1}
7	128	-3.1292×10^{-1}	-3.1292×10^{-1}	-3.1292×10^{-1}	-3.1292×10^{-1}	-3.1292×10^{-1}	-3.1292×10^{-1}	-3.1292×10^{-1}	-3.1292×10^{-1}
8	256	-3.2174×10^{-1}	-3.2174×10^{-1}	-3.2174×10^{-1}	-3.2174×10^{-1}	-3.2174×10^{-1}	-3.2174×10^{-1}	-3.2174×10^{-1}	-3.2174×10^{-1}

Table B.24: *The numerical orders of convergence to zero with respect to L for S_4, denoted by β_4, using the Preissman Box scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$, by considering $L = 2^\delta$ for $\delta = 0, \ldots, 9$ and fixed $N_x = 3^7$ and calculating β_4 similarly to Equation B.2. The values of L listed in the table are the smaller of the two values of L required to generate β_4. The other is identified by multiplying the listed value for L by two.*

Lax-Wendroff Scheme: $\beta_4 = \log\left(\frac{E_{4,2L}}{E_{4,L}}\right) / \log(2)$									
δ	$L = 2^\delta$ \ r	0	1	2	3	4	5	6	7
0	1	-7.6739×10^{-1}	-7.6739×10^{-1}	-7.6739×10^{-1}	-7.6739×10^{-1}	-7.6739×10^{-1}	-7.6739×10^{-1}	-7.6739×10^{-1}	-7.6739×10^{-1}
1	2	3.9801×10^{-1}	3.9801×10^{-1}	3.9801×10^{-1}	3.9801×10^{-1}	3.9801×10^{-1}	3.9801×10^{-1}	3.9801×10^{-1}	3.9801×10^{-1}
2	4	1.7561×10^{-1}	1.7561×10^{-1}	1.7561×10^{-1}	1.7561×10^{-1}	1.7561×10^{-1}	1.7561×10^{-1}	1.7561×10^{-1}	1.7561×10^{-1}
3	8	1.9694×10^{-2}	1.9694×10^{-2}	1.9694×10^{-2}	1.9694×10^{-2}	1.9694×10^{-2}	1.9694×10^{-2}	1.9694×10^{-2}	1.9694×10^{-2}
4	16	-6.5300×10^{-2}	-6.5300×10^{-2}	-6.5300×10^{-2}	-6.5300×10^{-2}	-6.5300×10^{-2}	-6.5300×10^{-2}	-6.5300×10^{-2}	-6.5300×10^{-2}
5	32	-9.9680×10^{-2}	-9.9680×10^{-2}	-9.9680×10^{-2}	-9.9680×10^{-2}	-9.9680×10^{-2}	-9.9680×10^{-2}	-9.9680×10^{-2}	-9.9680×10^{-2}
6	64	-1.0613×10^{-1}	-1.0613×10^{-1}	-1.0613×10^{-1}	-1.0613×10^{-1}	-1.0613×10^{-1}	-1.0613×10^{-1}	-1.0613×10^{-1}	-1.0613×10^{-1}
7	128	-9.9724×10^{-2}	-9.9724×10^{-2}	-9.9724×10^{-2}	-9.9724×10^{-2}	-9.9724×10^{-2}	-9.9724×10^{-2}	-9.9724×10^{-2}	-9.9724×10^{-2}
8	256	-8.8338×10^{-2}	-8.8338×10^{-2}	-8.8338×10^{-2}	-8.8338×10^{-2}	-8.8338×10^{-2}	-8.8338×10^{-2}	-8.8338×10^{-2}	-8.8338×10^{-2}

Table B.25: *The numerical orders of convergence to zero with respect to L for S_4, denoted by β_4, using the Lax-Wendroff scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$, by considering $L = 2^\delta$ for $\delta = 0, \ldots, 9$ and fixed $N_x = 3^7$ and calculating β_4 similarly to Equation B.2. The values of L listed in the table are the smaller of the two values of L required to generate β_4. The other is identified by multiplying the listed value for L by two.*

MNIMC Scheme: $\beta_4 = \log\left(\frac{E_{4,2L}}{E_{4,L}}\right) / \log(2)$									
δ	$L = 2^\delta$ \ r	0	1	2	3	4	5	6	7
0	1	-1.1699	-1.1699	-1.1699	-1.1699	-1.1699	-1.1699	-1.1699	-1.1699
1	2	5.2607×10^{-1}	5.2607×10^{-1}	5.2607×10^{-1}	5.2607×10^{-1}	5.2607×10^{-1}	5.2607×10^{-1}	5.2607×10^{-1}	5.2607×10^{-1}
2	4	3.0401×10^{-1}	3.0401×10^{-1}	3.0401×10^{-1}	3.0401×10^{-1}	3.0401×10^{-1}	3.0401×10^{-1}	3.0401×10^{-1}	3.0401×10^{-1}
3	8	1.6492×10^{-1}	1.6492×10^{-1}	1.6492×10^{-1}	1.6492×10^{-1}	1.6492×10^{-1}	1.6492×10^{-1}	1.6492×10^{-1}	1.6492×10^{-1}
4	16	8.6137×10^{-2}	8.6137×10^{-2}	8.6137×10^{-2}	8.6137×10^{-2}	8.6137×10^{-2}	8.6137×10^{-2}	8.6137×10^{-2}	8.6137×10^{-2}
5	32	4.4053×10^{-2}	4.4053×10^{-2}	4.4053×10^{-2}	4.4053×10^{-2}	4.4053×10^{-2}	4.4053×10^{-2}	4.4053×10^{-2}	4.4053×10^{-2}
6	64	2.2281×10^{-2}	2.2281×10^{-2}	2.2281×10^{-2}	2.2281×10^{-2}	2.2281×10^{-2}	2.2281×10^{-2}	2.2281×10^{-2}	2.2281×10^{-2}
7	128	1.1205×10^{-2}	1.1205×10^{-2}	1.1205×10^{-2}	1.1205×10^{-2}	1.1205×10^{-2}	1.1205×10^{-2}	1.1205×10^{-2}	1.1205×10^{-2}
8	256	5.6191×10^{-3}	5.6191×10^{-3}	5.6191×10^{-3}	5.6191×10^{-3}	5.6191×10^{-3}	5.6191×10^{-3}	5.6191×10^{-3}	5.6191×10^{-3}

Table B.26: *The numerical orders of convergence to zero with respect to L for S_4, denoted by β_4, using the MNIMC scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$, by considering $L = 2^\delta$ for $\delta = 0, \ldots, 9$ and fixed $N_x = 3^7$ and calculating β_4 similarly to Equation B.2. The values of L listed in the table are the smaller of the two values of L required to generate β_4. The other is identified by multiplying the listed value for L by two.*

B.4.3 analytically for the Upwind scheme

The analytical order of convergence of S_4, for the Upwind scheme, is calculated in the following.

$$\begin{aligned} E_4 &= |S_4 - 0| \\ &= \xi_1^2 + 2 \sum_{p=2}^{\frac{N_x+1}{2}} \xi_p^2 \\ &\leq K_2^2 + 2 \sum_{p=1}^{\frac{N_x-1}{2}} K_2^2 \\ &= K_2^2 N_x \end{aligned} \tag{B.16}$$

$$\Rightarrow \quad \frac{4D_3^2}{N_x^{2r+1}} \left(\xi_1^2 + 2 \sum_{p=2}^{\frac{N_x+1}{2}} \xi_p^2 \right) \leq 4D_3^2 K_2^2 N_x^{-2r} \tag{B.17}$$

$$\Rightarrow \quad \frac{4D_3^2}{N_x^{2r+1}} \left(\xi_1^2 + 2 \sum_{p=2}^{\frac{N_x+1}{2}} \xi_p^2 \right) = \mathcal{O}(N_x^{-2r}) \tag{B.18}$$

B.5 The orders of convergence for S_5...

B.5.1 with respect to N_x

Upwind Scheme: $\alpha_5 = \log\left(\frac{E_{5,3N_x}}{E_{5,N_x}}\right) / \log(3)$									
γ	$N_x = 3^\gamma$ \ r	0	1	2	3	4	5	6	7
2	9	2.0819×10^{-4}	−1.9998	−3.9998	−5.9998	−7.9998	−9.9998	−11.9998	−13.9998
3	27	-7.2752×10^{-10}	−2.0000	−4.0000	−6.0000	−8.0000	−10.0000	−12.0000	−14.0000
4	81	-1.0894×10^{-13}	−2.0000	−4.0000	−6.0000	−8.0000	−10.0000	−12.0000	−14.0000
5	243	-4.0423×10^{-16}	−2.0000	−4.0000	−6.0000	−8.0000	−10.0000	−12.0000	−14.0000
6	729	1.8190×10^{-15}	−2.0000	−4.0000	−6.0000	−8.0000	−10.0000	−12.0000	−14.0000

Table B.27: *The numerical orders of convergence to zero with respect to N_x for S_5, denoted by α_5, using the Upwind scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$, by considering $N_x = 3^\gamma$ for $\gamma = 2, \ldots, 7$ and fixed $L = 4$ and calculating them through α_5 as in Equation B.2. The values of N_x listed in the table are the smaller of the two values of N_x required to generate α_5. The other is identified by multiplying the listed value for N_x by three.*

Preissman Box Scheme: $\alpha_5 = \log\left(\frac{E_{5,3N_x}}{E_{5,N_x}}\right)/\log(3)$									
γ	$N_x = 3^\gamma$ \ r	0	1	2	3	4	5	6	7
2	9	-2.4089×10^{-2}	-2.0241	-4.0241	-6.0241	-8.0241	-10.0241	-12.0241	-14.0241
3	27	-2.2226×10^{-3}	-2.0022	-4.0022	-6.0022	-8.0022	-10.0022	-12.0022	-14.0022
4	81	-2.5568×10^{-4}	-2.0003	-4.0003	-6.0003	-8.0003	-10.0003	-12.0003	-14.0003
5	243	-2.8501×10^{-5}	-2.0000	-4.0000	-6.0000	-8.0000	-10.0000	-12.0000	-14.0000
6	729	-3.1679×10^{-6}	-2.0000	-4.0000	-6.0000	-8.0000	-10.0000	-12.0000	-14.0000

Table B.28: *The numerical orders of convergence to zero with respect to N_x for S_5, denoted by α_5, using the Preissman Box scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$, by considering $N_x = 3^\gamma$ for $\gamma = 2, \ldots, 7$ and fixed $L = 4$ and calculating them through α_5 as in Equation B.2. The values of N_x listed in the table are the smaller of the two values of N_x required to generate α_5. The other is identified by multiplying the listed value for N_x by three.*

Lax-Wendroff Scheme: $\alpha_5 = \log\left(\frac{E_{5,3N_x}}{E_{5,N_x}}\right)/\log(3)$									
γ	$N_x = 3^\gamma$ \ r	0	1	2	3	4	5	6	7
2	9	-1.3857×10^{-4}	-2.0001	-4.0001	-6.0001	-8.0001	-10.0001	-12.0001	-14.0001
3	27	-5.6348×10^{-5}	-2.0001	-4.0001	-6.0001	-8.0001	-10.0001	-12.0001	-14.0001
4	81	-4.4746×10^{-6}	-2.0000	-4.0000	-6.0000	-8.0000	-10.0000	-12.0000	-14.0000
5	243	-1.7418×10^{-6}	-2.0000	-4.0000	-6.0000	-8.0000	-10.0000	-12.0000	-14.0000
6	729	-3.5047×10^{-7}	-2.0000	-4.0000	-6.0000	-8.0000	-10.0000	-12.0000	-14.0000

Table B.29: *The numerical orders of convergence to zero with respect to N_x for S_5, denoted by α_5, using the Lax-Wendroff scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$, by considering $N_x = 3^\gamma$ for $\gamma = 2, \ldots, 7$ and fixed $L = 4$ and calculating them through α_5 as in Equation B.2. The values of N_x listed in the table are the smaller of the two values of N_x required to generate α_5. The other is identified by multiplying the listed value for N_x by three.*

B.5.2 with respect to L

Upwind Scheme: $\beta_5 = \log\left(\frac{E_{5,2L}}{E_{5,L}}\right)/\log(2)$									
δ	$L = 2^\delta$ \ r	0	1	2	3	4	5	6	7
0	1	6.0762×10^{-1}	6.0762×10^{-1}	6.0762×10^{-1}	6.0762×10^{-1}	6.0762×10^{-1}	6.0762×10^{-1}	6.0762×10^{-1}	6.0762×10^{-1}
1	2	9.5911×10^{-1}	9.5911×10^{-1}	9.5911×10^{-1}	9.5911×10^{-1}	9.5911×10^{-1}	9.5911×10^{-1}	9.5911×10^{-1}	9.5911×10^{-1}
2	4	6.9230×10^{-1}	6.9230×10^{-1}	6.9230×10^{-1}	6.9230×10^{-1}	6.9230×10^{-1}	6.9230×10^{-1}	6.9230×10^{-1}	6.9230×10^{-1}
3	8	4.7478×10^{-1}	4.7478×10^{-1}	4.7478×10^{-1}	4.7478×10^{-1}	4.7478×10^{-1}	4.7478×10^{-1}	4.7478×10^{-1}	4.7478×10^{-1}
4	16	3.1683×10^{-1}	3.1683×10^{-1}	3.1683×10^{-1}	3.1683×10^{-1}	3.1683×10^{-1}	3.1683×10^{-1}	3.1683×10^{-1}	3.1683×10^{-1}
5	32	2.0997×10^{-1}	2.0997×10^{-1}	2.0997×10^{-1}	2.0997×10^{-1}	2.0997×10^{-1}	2.0997×10^{-1}	2.0997×10^{-1}	2.0997×10^{-1}
6	64	1.3988×10^{-1}	1.3988×10^{-1}	1.3988×10^{-1}	1.3988×10^{-1}	1.3988×10^{-1}	1.3988×10^{-1}	1.3988×10^{-1}	1.3988×10^{-1}
7	128	9.4130×10^{-2}	9.4130×10^{-2}	9.4130×10^{-2}	9.4130×10^{-2}	9.4130×10^{-2}	9.4130×10^{-2}	9.4130×10^{-2}	9.4130×10^{-2}
8	256	6.4024×10^{-2}	6.4024×10^{-2}	6.4024×10^{-2}	6.4024×10^{-2}	6.4024×10^{-2}	6.4024×10^{-2}	6.4024×10^{-2}	6.4024×10^{-2}

Table B.30: *The numerical orders of convergence to zero with respect to L for S_5, denoted by β_5, using the Upwind scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$, by considering $L = 2^\delta$ for $\delta = 0, \ldots, 9$ and fixed $N_x = 3^7$ and calculating β_5 similarly to Equation B.2. The values of L listed in the table are the smaller of the two values of L required to generate β_5. The other is identified by multiplying the listed value for L by two.*

Preissman Box Scheme: $\beta_5 = \log\left(\frac{E_{5,2L}}{E_{5,L}}\right) / \log(2)$									
δ	$L = 2^\delta$ \ r	0	1	2	3	4	5	6	7
0	1	2.5805×10^{-1}	2.5805×10^{-1}	2.5805×10^{-1}	2.5805×10^{-1}	2.5805×10^{-1}	2.5805×10^{-1}	2.5805×10^{-1}	2.5805×10^{-1}
1	16	2.4430×10^{-1}	2.4430×10^{-1}	2.4430×10^{-1}	2.4430×10^{-1}	2.4430×10^{-1}	2.4430×10^{-1}	2.4430×10^{-1}	2.4430×10^{-1}
2	4	1.7456×10^{-2}	1.7456×10^{-2}	1.7456×10^{-2}	1.7456×10^{-2}	1.7456×10^{-2}	1.7456×10^{-2}	1.7456×10^{-2}	1.7456×10^{-2}
3	8	-1.0077×10^{-1}	-1.0077×10^{-1}	-1.0077×10^{-1}	-1.0077×10^{-1}	-1.0077×10^{-1}	-1.0077×10^{-1}	-1.0077×10^{-1}	-1.0077×10^{-1}
4	16	-1.8096×10^{-1}	-1.8096×10^{-1}	-1.8096×10^{-1}	-1.8096×10^{-1}	-1.8096×10^{-1}	-1.8096×10^{-1}	-1.8096×10^{-1}	-1.8096×10^{-1}
5	32	-2.3319×10^{-1}	-2.3319×10^{-1}	-2.3319×10^{-1}	-2.3319×10^{-1}	-2.3319×10^{-1}	-2.3319×10^{-1}	-2.3319×10^{-1}	-2.3319×10^{-1}
6	64	-2.6700×10^{-1}	-2.6700×10^{-1}	-2.6700×10^{-1}	-2.6700×10^{-1}	-2.6700×10^{-1}	-2.6700×10^{-1}	-2.6700×10^{-1}	-2.6700×10^{-1}
7	128	-2.8902×10^{-1}	-2.8902×10^{-1}	-2.8902×10^{-1}	-2.8902×10^{-1}	-2.8902×10^{-1}	-2.8902×10^{-1}	-2.8902×10^{-1}	-2.8902×10^{-1}
8	256	-3.0357×10^{-1}	-3.0357×10^{-1}	-3.0357×10^{-1}	-3.0357×10^{-1}	-3.0357×10^{-1}	-3.0357×10^{-1}	-3.0357×10^{-1}	-3.0357×10^{-1}

Table B.31: *The numerical orders of convergence to zero with respect to L for S_5, denoted by β_5, using the Preissman Box scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$, by considering $L = 2^\delta$ for $\delta = 0, \ldots, 9$ and fixed $N_x = 3^7$ and calculating β_5 similarly to Equation B.2. The values of L listed in the table are the smaller of the two values of L required to generate β_5. The other is identified by multiplying the listed value for L by two.*

Lax-Wendroff Scheme: $\beta_5 = \log\left(\frac{E_{5,2L}}{E_{5,L}}\right) / \log(2)$									
δ	$L = 2^\delta$ \ r	0	1	2	3	4	5	6	7
0	1	5.4338×10^{-1}	5.4338×10^{-1}	5.4338×10^{-1}	5.4338×10^{-1}	5.4338×10^{-1}	5.4338×10^{-1}	5.4338×10^{-1}	5.4338×10^{-1}
1	2	6.1676×10^{-1}	6.1676×10^{-1}	6.1676×10^{-1}	6.1676×10^{-1}	6.1676×10^{-1}	6.1676×10^{-1}	6.1676×10^{-1}	6.1676×10^{-1}
2	4	3.9793×10^{-1}	3.9793×10^{-1}	3.9793×10^{-1}	3.9793×10^{-1}	3.9793×10^{-1}	3.9793×10^{-1}	3.9793×10^{-1}	3.9793×10^{-1}
3	8	2.3110×10^{-1}	2.3110×10^{-1}	2.3110×10^{-1}	2.3110×10^{-1}	2.3110×10^{-1}	2.3110×10^{-1}	2.3110×10^{-1}	2.3110×10^{-1}
4	16	1.2020×10^{-1}	1.2020×10^{-1}	1.2020×10^{-1}	1.2020×10^{-1}	1.2020×10^{-1}	1.2020×10^{-1}	1.2020×10^{-1}	1.2020×10^{-1}
5	32	5.6728×10^{-2}	5.6728×10^{-2}	5.6728×10^{-2}	5.6728×10^{-2}	5.6728×10^{-2}	5.6728×10^{-2}	5.6728×10^{-2}	5.6728×10^{-2}
6	64	2.2837×10^{-2}	2.2837×10^{-2}	2.2837×10^{-2}	2.2837×10^{-2}	2.2837×10^{-2}	2.2837×10^{-2}	2.2837×10^{-2}	2.2837×10^{-2}
7	128	5.1822×10^{-3}	5.1822×10^{-3}	5.1822×10^{-3}	5.1822×10^{-3}	5.1822×10^{-3}	5.1822×10^{-3}	5.1822×10^{-3}	5.1822×10^{-3}
8	256	-3.8556×10^{-3}	-3.8556×10^{-3}	-3.8556×10^{-3}	-3.8556×10^{-3}	-3.8556×10^{-3}	-3.8556×10^{-3}	-3.8556×10^{-3}	-3.8556×10^{-3}

Table B.32: *The numerical orders of convergence to zero with respect to L for S_5, denoted by β_5, using the Lax-Wendroff scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$, by considering $L = 2^\delta$ for $\delta = 0, \ldots, 9$ and fixed $N_x = 3^7$ and calculating β_5 similarly to Equation B.2. The values of L listed in the table are the smaller of the two values of L required to generate β_5. The other is identified by multiplying the listed value for L by two.*

B.5.3 analytically for the Upwind scheme

The analytical order of convergence of S_5, for the Upwind scheme, is calculated in the following.

$$\begin{aligned}
E_5 &= |S_5 - 0| \\
&= \sum_{p=2}^{\frac{N_x+1}{2}} |1-\nu_p|\xi_p \\
&\leq \frac{K_1 K_2}{N_x^2} \sum_{p=1}^{\frac{N_x-1}{2}} p^2 \\
&= \frac{K_1 K_2}{N_x^2} \left(\frac{N_x-1}{2}\right) \left[2\left(\frac{N_x-1}{2}\right)^2 + 3\left(\frac{N_x-1}{2}\right) + 1\right] \\
&\quad \text{where } \sum_{p=1}^{\frac{N_x-1}{2}} p^2 \text{ is calculated using [97],} \\
&< \frac{K_1 K_2}{24} N_x \qquad \text{(B.19)}
\end{aligned}$$

$$\Rightarrow \quad \frac{8D_3^2}{N_x^{2r+1}} \sum_{p=2}^{\frac{N_x+1}{2}} |1-\nu_p|\xi_p < \frac{8D_3^2 K_1 K_2}{24} N_x^{-2r} \qquad \text{(B.20)}$$

$$\Rightarrow \quad \frac{8D_3^2}{N_x^{2r+1}} \sum_{p=2}^{\frac{N_x+1}{2}} |1-\nu_p|\xi_p = \mathcal{O}(N_x^{-2r}) \qquad \text{(B.21)}$$

(B.22)

B.6 The orders of convergence for S_6...

B.6.1 with respect to N_x

Upwind Scheme: $\alpha_6 = \log\left(\frac{E_{6,3N_x}}{E_{6,N_x}}\right) / \log(3)$									
γ	r / $N_x = 3^\gamma$	0	1	2	3	4	5	6	7
2	9	1.3752×10^{-2}	-1.8946	-3.5670	-4.9046	-6.0224	-7.0572	-8.0673	-9.0703
3	27	1.5019×10^{-3}	-1.9675	-3.7095	-4.9764	-6.0116	-7.0122	-8.0120	-9.0104
4	81	1.6673×10^{-4}	-1.9894	-3.7803	-4.9930	-6.0023	-7.0016	-8.0013	-9.0012
5	243	1.8524×10^{-5}	-1.9965	-3.8231	-4.9977	-6.0004	-7.0002	-8.0001	-9.0001
6	729	2.0581×10^{-6}	-1.9988	-3.8520	-4.9993	-6.0001	-7.0000	-8.0000	-9.0000

Table B.33: *The numerical orders of convergence to zero with respect to N_x for S_6, denoted by α_6, using the Upwind scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$, by considering $N_x = 3^\gamma$ for $\gamma = 2, \ldots, 7$ and fixed $L = 4$ and calculating them through α_6 as in Equation B.2. The values of N_x listed in the table are the smaller of the two values of N_x required to generate α_6. The other is identified by multiplying the listed value for N_x by three.*

Preissman Box Scheme: $\alpha_6 = \log\left(\frac{E_{6,3N_x}}{E_{6,N_x}}\right) / \log(3)$									
γ	$N_x = 3^\gamma$ \ r	0	1	2	3	4	5	6	7
2	9	-1.3691×10^{-2}	-1.9984	-3.9234	-5.6121	-6.9253	-8.0088	-9.0243	-10.0260
3	27	-1.4505	-1.9999	-3.9755	-5.7291	-6.9781	-8.0035	-9.0037	-10.0032
4	81	-1.6758×10^{-4}	-2.0000	-3.9919	-5.7915	-6.9930	-8.0007	-9.0005	-10.0004
5	243	-1.8688×10^{-5}	-2.0000	-3.9973	-5.8304	-6.9977	-8.0001	-9.0001	-10.0000
6	729	-2.0772×10^{-6}	-2.0000	-3.9991	-5.8571	-6.9992	-8.0000	-9.0000	-10.0000

Table B.34: *The numerical orders of convergence to zero with respect to N_x for S_6, denoted by α_6, using the Preissman Box scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$, by considering $N_x = 3^\gamma$ for $\gamma = 2, \ldots, 7$ and fixed $L = 4$ and calculating them through α_6 as in Equation B.2. The values of N_x listed in the table are the smaller of the two values of N_x required to generate α_6. The other is identified by multiplying the listed value for N_x by three.*

Lax-Wendroff Scheme: $\alpha_6 = \log\left(\frac{E_{6,3N_x}}{E_{6,N_x}}\right)/\log(3)$									
γ	$N_x = 3^\gamma$ \ r	0	1	2	3	4	5	6	7
2	9	-1.0031×10^{-3}	-1.9897	-3.9012	-5.5605	-6.8645	-7.9566	-8.9797	-9.9853
3	27	-1.3215×10^{-4}	-1.9989	-3.9691	-5.7043	-6.9574	-7.9932	-8.9970	-9.9975
4	81	1.0629×10^{-4}	-1.9920	-3.6732	-3.8326	-4.0881	-4.8323	-5.7403	-6.7024
5	243	-8.2311×10^{-5}	-2.0052	-4.1953	-6.7511	-8.7376	-10.4637	-11.9331	-13.1701
6	729	-2.1736×10^{-5}	-2.0015	-4.0634	-6.3613	-7.8791	-8.6563	-9.3204	-10.1265
7	6561	-9.5495×10^{-6}	-2.0006	-4.0274	-6.1387	-7.2276	-8.0442	-9.0057	-10.0007
8	19683	-1.6864×10^{-6}	-2.0001	-4.0053	-5.9567	-7.0349	-8.0022	-9.0001	-10.0000
9	59049	8.0605×10^{-7}	-2.0000	-3.9978	-5.9010	-7.0075	-8.0002	-9.0000	-10.0000
10	177147	-6.4607×10^{-7}	-2.0000	-4.0016	-9.9339	-7.0031	-8.0000	-9.0000	-10.0000
11	531441	-5.8297×10^{-8}	-2.0000	-4.0002	-5.9305	-7.0009	-8.0000	-9.0000	-10.0000

Table B.35: *The numerical orders of convergence to zero with respect to N_x for S_6, denoted by α_6, using the Lax-Wendroff scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$, by considering $N_x = 3^\gamma$ for $\gamma = 2, \ldots, 12$ and fixed $L = 4$ and calculating them through α_6 as in Equation B.2. The values of N_x listed in the table are the smaller of the two values of N_x required to generate α_6. The other is identified by multiplying the listed value for N_x by three.*

B.6.2 with respect to L

Upwind Scheme: $\beta_6 = \log\left(\frac{E_{6,2L}}{E_{6,L}}\right)/\log(2)$									
δ	$L = 2^\delta$ \ r	0	1	2	3	4	5	6	7
0	1	6.0437×10^{-1}	5.6352×10^{-1}	4.5161×10^{-1}	4.1555×10^{-1}	4.1505×10^{-1}	4.1504×10^{-1}	4.1504×10^{-1}	4.1504×10^{-1}
1	2	1.0736	1.1880	1.2614	1.2633	1.2630	1.2630	1.2630	1.2630
2	4	8.4668×10^{-1}	1.0146	1.1427	1.1523	1.1520	1.1520	1.1520	1.1520
3	8	6.5663×10^{-1}	8.7324×10^{-1}	1.0617	1.0829	1.0825	1.0825	1.0825	1.0825
4	16	5.1078×10^{-1}	7.6542×10^{-1}	1.0092	1.0436	1.0431	1.0431	1.0431	1.0431
5	32	4.0427×10^{-1}	6.8675×10^{-1}	9.7557×10^{-1}	1.0227	1.0221	1.0221	1.0221	1.0221
6	64	3.2767×10^{-1}	6.3075×10^{-1}	9.5356×10^{-1}	1.0120	1.0113	1.0112	1.0112	1.0112
7	128	2.7220×10^{-1}	5.9130×10^{-1}	9.3855×10^{-1}	1.0068	1.0060	1.0058	1.0057	1.0058
8	256	2.3124×10^{-1}	5.6353×10^{-1}	9.2757×10^{-1}	1.0044	1.0034	1.0031	1.0030	1.0030

Table B.36: *The numerical orders of convergence to zero with respect to L for S_6, denoted by β_6, using the Upwind scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$, by considering $L = 2^\delta$ for $\delta = 0, \ldots, 9$ and fixed $N_x = 3^7$ and calculating β_6 similarly to Equation B.2. The values of L listed in the table are the smaller of the two values of L required to generate β_6. The other is identified by multiplying the listed value for L by two.*

Preissman Box Scheme: $\beta_6 = \log\left(\frac{E_{6,2L}}{E_{6,L}}\right) / \log(2)$									
δ	$L = 2^\delta$ \ r	0	1	2	3	4	5	6	7
0	1	2.8339×10^{-1}	3.1445×10^{-1}	3.5461×10^{-1}	4.0463×10^{-1}	4.1498×10^{-1}	4.1504×10^{-1}	4.1504×10^{-1}	4.1504×10^{-1}
1	16	4.4862×10^{-1}	6.7438×10^{-1}	9.3277×10^{-1}	1.2104	1.2628	1.2630	1.2630	1.2630
2	4	2.5233×10^{-1}	5.1451×10^{-1}	8.0716×10^{-1}	1.0990	1.1517	1.1520	1.1520	1.1520
3	8	1.5863×10^{-1}	4.4136×10^{-1}	7.4545×10^{-1}	1.0314	1.0821	1.0825	1.0825	1.0825
4	16	9.7593×10^{-2}	3.9577×10^{-1}	7.0874×10^{-1}	9.9213×10^{-1}	1.0427	1.0431	1.0431	1.0431
5	32	5.9780×10^{-2}	3.6900×10^{-1}	6.8840×10^{-1}	9.7035×10^{-1}	1.0215	1.0220	1.0220	1.0220
6	64	3.6718×10^{-2}	3.5368×10^{-1}	6.7755×10^{-1}	9.5823×10^{-1}	1.0105	1.0111	1.0111	1.0111
7	128	2.2675×10^{-2}	3.4499×10^{-1}	6.7185×10^{-1}	9.5111×10^{-1}	1.0048	1.0056	1.0056	1.0056
8	256	1.4064×10^{-2}	3.4005×10^{-1}	6.6883×10^{-1}	9.4645×10^{-1}	1.0018	1.0028	1.0028	1.0028

Table B.37: *The numerical orders of convergence to zero with respect to L for S_6, denoted by β_6, using the Preissman Box scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$, by considering $L = 2^\delta$ for $\delta = 0, \ldots, 9$ and fixed $N_x = 3^7$ and calculating β_6 similarly to Equation B.2. The values of L listed in the table are the smaller of the two values of L required to generate β_6. The other is identified by multiplying the listed value for L by two.*

Lax-Wendroff Scheme: $\beta_6 = \log\left(\frac{E_{6,2L}}{E_{6,L}}\right) / \log(2)$									
δ	r / $L = 2^\delta$	0	1	2	3	4	5	6	7
0	1	5.4550×10^{-1}	5.2999×10^{-1}	3.0833×10^{-1}	-3.6003×10^{-1}	-1.7169×10^{-1}	2.5421×10^{-1}	3.9275×10^{-1}	4.1247×10^{-1}
1	2	7.7337×10^{-1}	9.3844×10^{-1}	1.0292	6.5200×10^{-1}	8.6225×10^{-1}	1.1772	1.2519	1.2617
2	4	5.6169×10^{-1}	7.4822×10^{-1}	9.1399×10^{-1}	7.7571×10^{-1}	9.4415×10^{-1}	1.1147	1.1474	1.1515
3	8	3.9222×10^{-1}	5.9475×10^{-1}	8.0896×10^{-1}	8.2723×10^{-1}	9.6377×10^{-1}	1.0630	1.0801	1.0822
4	16	2.6874×10^{-1}	4.8085×10^{-1}	7.3317×10^{-1}	8.7703×10^{-1}	9.8190×10^{-1}	1.0336	1.0419	1.0429
5	32	1.8715×10^{-1}	4.0368×10^{-1}	6.8403×10^{-1}	9.0816×10^{-1}	9.9221×10^{-1}	1.0176	1.0215	1.0220
6	64	1.3460×10^{-1}	3.5376×10^{-1}	6.5545×10^{-1}	9.2543×10^{-1}	9.9753×10^{-1}	1.0092	1.0109	1.0111
7	128	1.0011×10^{-1}	3.2217×10^{-1}	6.4081×10^{-1}	9.3370×10^{-1}	9.9980×10^{-1}	1.0049	1.0055	1.0056
8	256	7.6528×10^{-2}	3.0256×10^{-1}	6.3466×10^{-1}	9.3594×10^{-1}	1.0001	1.0026	1.0028	1.0028

Table B.38: *The numerical orders of convergence to zero with respect to L for S_6, denoted by β_6, using the Lax-Wendroff scheme. The numerical results are generated using initial condition regularities $r = 0, \ldots, 7$, by considering $L = 2^\delta$ for $\delta = 0, \ldots, 9$ and fixed $N_x = 3^7$ and calculating β_6 similarly to Equation B.2. The values of L listed in the table are the smaller of the two values of L required to generate β_6. The other is identified by multiplying the listed value for L by two.*

B.6.3 analytically for the Upwind scheme

The analytical order of convergence of S_6, for the Upwind scheme, is calculated in the following.

$$\begin{aligned}
E_6 &= |S_6 - 0| \\
&= \sum_{p=2}^{\frac{N_x+1}{2}} \frac{|1-\nu_p|\xi_p}{(p-1)^{r+1}} \\
&\leq \frac{K_1 K_2 L}{N_x^2} \sum_{p=1}^{\frac{N_x-1}{2}} p^{1-r} \\
&\quad \text{where } \sum_{p=1}^{\frac{N_x-1}{2}} p^{1-r} \text{ is calculated using [97] for } r = 0, 1, \\
&< \begin{cases} \frac{3K_1K_2}{8} L N_x^0, & \text{for } r = 0, \\ \frac{K_1K_2}{2} L N_x^{-1}, & \text{for } r = 1, \\ K_1 K_2 L N_x^{-2} \left(\frac{1}{N_x - 1} + \log(N_x - 1) - \log(2) + \gamma \right), & \text{for } r = 2, \\ K_1 K_2 L N_x^{-2} \zeta(r-1), & \text{for } r \geq 3. \end{cases} \qquad \text{(B.23)} \\
&\quad \text{where the case for } r = 2 \text{ is due to (B.8).}
\end{aligned}$$

$$\Rightarrow \quad \frac{4D_2D_3L}{N_x^r} \sum_{p=2}^{\frac{N_x+1}{2}} \frac{|1-\nu_p|\xi_p}{(p-1)^{r+1}} < \begin{cases} \frac{3D_2D_3K_1K_2L}{4} N_x^0, & \text{for } r = 0, \\ 2D_2D_3K_1K_2LN_x^{-2}, & \text{for } r = 1, \\ 4D_2D_3K_1K_2(2+\gamma)LN_x^{-4}\log(N_x), & \text{for } r = 2, \\ 4D_2D_3K_1K_2\zeta(r-1)LN_x^{-r-2}, & \text{for } r \geq 3, \end{cases} \qquad \text{(B.24)}$$

where the case for $r = 2$ is due to (B.9),

$$\Rightarrow \quad \frac{4D_2D_3L}{N_x^r} \sum_{p=2}^{\frac{N_x+1}{2}} \frac{|1-\nu_p|\xi_p}{(p-1)^{r+1}} < \begin{cases} \mathcal{O}(LN_x^{-2r}), & \text{for } r = 0, 1, \\ \mathcal{O}(LN_x^{-4}\log(N_x)), & \text{for } r = 2, \\ \mathcal{O}(LN_x^{-r-2}), & \text{for } r \geq 3. \end{cases} \qquad \text{(B.25)}$$

APPENDIX C

Numerical Orders of Convergence for the 2D Linear Advection Problem

In this chapter of the Appendix, we consider the numerical results for the 2D linear advection problem. Consider the summations R_1 to R_3 of Section 5.11. Let E_k denote the magnitude of the error in R_k when compared to zero, for $k = 1, \ldots, 3$. Assume the error has the following form when $N_x = N_y$,

$$E_k \approx C_k N_x^{\alpha_k} L^{\beta_k}. \tag{C.1}$$

Given this, the order of convergence with respect to both N_x and L can be calculated. In order to identify α_k, N_x is varied by the same factor whilst L remains constant. We require N_x to be odd due to the use of the MNIMC scheme, so N_x is chosen so that $N_x = 3^\gamma$ for $\gamma = 1, \ldots, 7$ for the Upwind and Crank-Nicolson schemes whilst $L = 4$. Let E_{k,N_x} denote the magnitude of the error between R_k and zero, as $N_x \to \infty$. Then,

$$\log\left(\frac{E_{k,3N_x}}{E_{k,N_x}}\right) / \log(3) = \log\left(\frac{C_k(3N_x)^{\alpha_k}}{C_k N_x^{\alpha_k}}\right) / \log(3) = \alpha_k, \tag{C.2}$$

calculates α_k the order of convergence of E_k with respect to N_x. A similar calculation can be performed to identify β_k. In this instance, L is chosen such that $L = 2^\delta$ for $\delta = 0, \ldots, 9$, whilst $N_x = 3^7$. The Tables in the following Sections present the numerical values for α_k for $k = 1, \ldots, 3$.

C.1 The orders of convergence for R_1...

C.1.1 with respect to N_xN_y for the Upwind scheme

Upwind Scheme, $r_2 = 0$: $\alpha_1 = \log\left(\frac{E_{1,3N_x}}{E_{1,N_x}}\right)/\log(3)$									
γ \ r_1	0	1	2	3	4	5	6	7	$r_1 \gg 1$
1	6.2723×10^1	6.2131×10^1	6.2034×10^1	6.2019×10^1	6.2017×10^1	6.2016×10^1	6.2016×10^1	6.2016×10^1	6.2016×10^1
2	1.6371	1.4296	1.4653	1.4703	1.4708	1.4709	1.4709	1.4709	1.4709
3	1.2819	1.1791	1.1874	1.1864	1.1861	1.1860	1.1860	1.1860	1.1860
4	1.1265	1.0712	1.0701	1.0693	1.0691	1.0691	1.0690	1.0690	1.0690
5	1.0551	1.0259	1.0245	1.0242	1.0241	1.0241	1.0241	1.0241	1.0241
6	1.0232	1.0090	1.0083	1.0082	1.0082	1.0082	1.0082	1.0082	1.0082

Table C.1: *The numerical orders of convergence to zero with respect to N_x when $N_x = N_y$ for R_1, denoted by α_1, using the Upwind scheme. The numerical results are generated using initial condition regularities $r_1 = 0, \ldots, 7$ and $r_1 \gg 1$ with $r_2 = 0$, by considering $N_x = 3^\gamma$ for $\gamma = 1, \ldots, 7$ and fixed $L = 4$ and calculating them through α_1 as in Equation (C.2). The values of γ listed in the table are the smaller of the two values of γ used to generate the N_x required to calculate α_1. These values were generated using $\mu_1 = \mu_2 = 1$ and $h_1 = h_2 = \frac{1}{2}$. Equation (5.120) was used to generate the order of convergence when $r_1 \gg 1$ and $r_2 = 0$.*

Upwind Scheme, $r_2 = 1$: $\alpha_1 = \log\left(\frac{E_{1,3N_x}}{E_{1,N_x}}\right) / \log(3)$									
γ \ r_1	0	1	2	3	4	5	6	7	$r_1 \gg 1$
1	6.2131×10^{1}	6.0627×10^{1}	6.0102×10^{1}	6.0005×10^{1}	5.9988×10^{1}	5.9985×10^{1}	5.9984×10^{1}	5.9984×10^{1}	5.9983×10^{1}
2	1.4296	7.9110×10^{-2}	-1.5000×10^{-2}	4.1945×10^{-2}	5.4592×10^{-2}	5.7210×10^{-2}	5.7796×10^{-2}	5.7935×10^{-2}	5.7980×10^{-2}
3	1.1791	-4.5349×10^{-1}	-4.9170×10^{-1}	-4.7063×10^{-1}	-4.6894×10^{-1}	-4.6874×10^{-1}	-4.6871×10^{-1}	-4.6870×10^{-1}	-4.6870×10^{-1}
4	1.0712	-7.3776×10^{-1}	-7.5425×10^{-1}	-7.4865×10^{-1}	-7.4861×10^{-1}	-7.4864×10^{-1}	-7.4865×10^{-1}	-7.4865×10^{-1}	-7.4865×10^{-1}
5	1.0259	-8.8005×10^{-1}	-8.8752×10^{-1}	-8.8628×10^{-1}	-8.8639×10^{-1}	-8.8643×10^{-1}	-8.8644×10^{-1}	-8.8644×10^{-1}	-8.8644×10^{-1}
6	1.0090	-9.4741×10^{-1}	-9.5087×10^{-1}	-9.5069×10^{-1}	-9.5077×10^{-1}	-9.5079×10^{-1}	-9.5080×10^{-1}	-9.5080×10^{-1}	-9.5080×10^{-1}

Table C.2: *The numerical orders of convergence to zero with respect to N_x when $N_x = N_y$ for R_1, denoted by α_1, using the Upwind scheme. The numerical results are generated using initial condition regularities $r_1 = 0, \ldots, 7$ and $r_1 \gg 1$ with $r_2 = 1$, by considering $N_x = 3^{\gamma}$ for $\gamma = 1, \ldots, 7$ and fixed $L = 4$ and calculating them through α_1 as in Equation (C.2). The values of γ listed in the table are the smaller of the two values of γ used to generate the N_x required to calculate α_1. These values were generated using $\mu_1 = \mu_2 = 1$ and $h_1 = h_2 = \frac{1}{2}$. Equation (5.120) was used to generate the order of convergence when $r_1 \gg 1$ and $r_2 = 1$.*

Upwind Scheme, $r_2 = 2$: $\alpha_1 = \log\left(\frac{E_{1,3N_x}}{E_{1,N_x}}\right)/\log(3)$									
γ \ r_1	0	1	2	3	4	5	6	7	$r_1 \gg 1$
1	6.2034×10^1	6.0102×10^1	5.8789×10^1	5.8299×10^1	5.8187×10^1	5.8163×10^1	5.8158×10^1	5.8157×10^1	5.8156×10^1
2	1.4653	-1.5000×10^{-2}	-1.0079	-1.0750	-1.0283	-1.0147	-1.0113	-1.0105	-1.0103
3	1.1874	-4.9170×10^{-1}	-1.5846	-1.6098	-1.5907	-1.5873	-1.5866	-1.5864	-1.5863
4	1.0701	-7.5425×10^{-1}	-1.8440	-1.8526	-1.8462	-1.8453	-1.8451	-1.8450	-1.8450
5	1.0245	-8.8752×10^{-1}	-1.9450	-1.9479	-1.9458	-1.9455	-1.9454	-1.9454	-1.9454
6	1.0083	-9.5082×10^{-1}	-1.9919	-2.0036	-2.0046	-2.0049	-2.0050	-2.0050	-2.0050

Table C.3: *The numerical orders of convergence to zero with respect to N_x when $N_x = N_y$ for R_1, denoted by α_1, using the Upwind scheme. The numerical results are generated using initial condition regularities $r_1 = 0, \ldots, 7$ and $r_1 \gg 1$ with $r_2 = 2$, by considering $N_x = 3^\gamma$ for $\gamma = 1, \ldots, 7$ and fixed $L = 4$ and calculating them through α_1 as in Equation (C.2). The values of γ listed in the table are the smaller of the two values of γ used to generate the N_x required to calculate α_1. These values were generated using $\mu_1 = \mu_2 = 1$ and $h_1 = h_2 = \frac{1}{2}$. Equation (5.120) was used to generate the order of convergence when $r_1 \gg 1$ and $r_2 = 2$.*

Upwind Scheme, $r_2 = 3$: $\alpha_1 = \log\left(\frac{E_{1,3N_x}}{E_{1,N_x}}\right)/\log(3)$									
γ \ r_1	0	1	2	3	4	5	6	7	$r_1 \gg 1$
1	6.2019×10^1	6.0005×10^1	5.8299×10^1	5.7171×10^1	5.6707×10^1	5.6580×10^1	5.6549×10^1	5.6542×10^1	5.6539×10^1
2	1.4703	4.1945×10^{-2}	-1.0750	-1.6019	-1.6377	-1.6139	-1.6052	-1.6029	-1.6021
3	1.1864	-4.7063×10^{-1}	-1.6098	-1.9362	-1.9432	-1.9381	-1.9367	-1.9364	-1.9363
4	1.0693	-7.4865×10^{-1}	-1.8526	-1.9937	-1.9944	-1.9938	-1.9937	-1.9937	-1.9937
5	1.0242	-8.8628×10^{-1}	-1.9479	-1.9995	-1.9995	-1.9995	-1.9995	-1.9995	-1.9995
6	1.0082	-9.5063×10^{-1}	-1.9914	-2.0276	-2.0369	-2.0402	-2.0411	-2.0414	-2.0414

Table C.4: *The numerical orders of convergence to zero with respect to N_x when $N_x = N_y$ for R_1, denoted by α_1, using the Upwind scheme. The numerical results are generated using initial condition regularities $r_1 = 0, \ldots, 7$ and $r_1 \gg 1$ with $r_2 = 3$, by considering $N_x = 3^\gamma$ for $\gamma = 1, \ldots, 7$ and fixed $L = 4$ and calculating them through α_1 as in Equation (C.2). The values of γ listed in the table are the smaller of the two values of γ used to generate the N_x required to calculate α_1. These values were generated using $\mu_1 = \mu_2 = 1$ and $h_1 = h_2 = \frac{1}{2}$. Equation (5.120) was used to generate the order of convergence when $r_1 \gg 1$ and $r_2 = 3$.*

Upwind Scheme, $r_2 = 4$: $\alpha_1 = \log\left(\frac{E_{1,3N_x}}{E_{1,N_x}}\right)/\log(3)$									
γ \ r_1	0	1	2	3	4	5	6	7	$r_1 \gg 1$
1	6.2017×10^1	5.9988×10^1	5.8187×10^1	5.6707×10^1	5.5717×10^1	5.5271×10^1	5.5134×10^1	5.5098×10^1	5.5086×10^1
2	1.4708	5.4592×10^{-2}	-1.0283	-1.6377	-1.8379	-1.8524	-1.8433	-1.8394	-1.8380
3	1.1861	-4.6894×10^{-1}	-1.5907	-1.9432	-1.9885	-1.9898	-1.9889	-1.9886	-1.9885
4	1.0691	-7.4861×10^{-1}	-1.8462	-1.9944	-1.9993	-1.9993	-1.9993	-1.9993	-1.9993
5	1.0241	-8.8639×10^{-1}	-1.9458	-1.9995	-1.9999	-1.9999	-1.9999	-1.9999	-1.9999
6	1.0082	-9.5071×10^{-1}	-1.9910	-2.0201	-2.0190	-2.0213	-2.0233	-2.0240	-2.0242

Table C.5: *The numerical orders of convergence to zero with respect to N_x when $N_x = N_y$ for R_1, denoted by α_1, using the Upwind scheme. The numerical results are generated using initial condition regularities $r_1 = 0, \ldots, 7$ and $r_1 \gg 1$ with $r_2 = 4$, by considering $N_x = 3^\gamma$ for $\gamma = 1, \ldots, 7$ and fixed $L = 4$ and calculating them through α_1 as in Equation (C.2). The values of γ listed in the table are the smaller of the two values of γ used to generate the N_x required to calculate α_1. These values were generated using $\mu_1 = \mu_2 = 1$ and $h_1 = h_2 = \frac{1}{2}$. Equation (5.120) was used to generate the order of convergence when $r_1 \gg 1$ and $r_2 = 4$.*

Upwind Scheme, $r_2 = 5$: $\alpha_1 = \log\left(\frac{E_{1,3N_x}}{E_{1,N_x}}\right)/\log(3)$									
γ \ r_1	0	1	2	3	4	5	6	7	$r_1 \gg 1$
1	6.2016×10^1	5.9985×10^1	5.8163×10^1	5.6580×10^1	5.5271×10^1	5.4362×10^1	5.3926×10^1	5.3783×10^1	5.3731×10^1
2	1.4709	5.7210×10^{-2}	−1.0147	−1.6139	−1.8524	−1.9194	−1.9249	−1.9215	−1.9194
3	1.1860	-4.6874×10^{-1}	−1.5873	−1.9381	−1.9898	−1.9958	−1.9961	−1.9959	−1.9958
4	1.0691	-7.4864×10^{-1}	−1.8453	−1.9938	−1.9993	−1.9997	−1.9997	−1.9997	−1.9997
5	1.0241	-8.8643×10^{-1}	−1.9455	−1.9995	−1.9999	−2.0000	−2.0000	−2.0000	−2.0000
6	1.0082	-9.5073×10^{-1}	−1.9911	−2.0202	−2.0141	−2.0093	−2.0095	−2.0104	−2.0109

Table C.6: *The numerical orders of convergence to zero with respect to N_x when $N_x = N_y$ for R_1, denoted by α_1, using the Upwind scheme. The numerical results are generated using initial condition regularities $r_1 = 0, \ldots, 7$ and $r_1 \gg 1$ with $r_2 = 5$, by considering $N_x = 3^\gamma$ for $\gamma = 1, \ldots, 7$ and fixed $L = 4$ and calculating them through α_1 as in Equation (C.2). The values of γ listed in the table are the smaller of the two values of γ used to generate the N_x required to calculate α_1. These values were generated using $\mu_1 = \mu_2 = 1$ and $h_1 = h_2 = \frac{1}{2}$. Equation (5.120) was used to generate the order of convergence when $r_1 \gg 1$ and $r_2 = 5$.*

Upwind Scheme, $r_2 = 6$: $\alpha_1 = \log\left(\frac{E_{1,3N_x}}{E_{1,N_x}}\right)/\log(3)$									
γ \ r_1	0	1	2	3	4	5	6	7	$r_1 \gg 1$
1	6.2016×10^{1}	5.9984×10^{1}	5.8158×10^{1}	5.6549×10^{1}	5.5134×10^{1}	5.3926×10^{1}	5.3057×10^{1}	5.2626×10^{1}	5.2427×10^{1}
2	1.4709	5.7796×10^{-2}	-1.0113	-1.6052	-1.8433	-1.9249	-1.9480	-1.9502	-1.9480
3	1.1860	-4.6871×10^{-1}	-1.5866	-1.9367	-1.9889	-1.9961	-1.9973	-1.9974	-1.9973
4	1.0690	-7.4865×10^{-1}	-1.8451	-1.9937	-1.9993	-1.9997	-1.9998	-1.9998	-1.9998
5	1.0241	-8.8644×10^{-1}	-1.9454	-1.9995	-1.9999	-2.0000	-2.0000	-2.0000	-2.0000
6	1.0082	-9.5073×10^{-1}	-1.9911	-2.0206	-2.0145	-2.0073	-2.0042	-2.0040	-2.0046

Table C.7: *The numerical orders of convergence to zero with respect to N_x when $N_x = N_y$ for R_1, denoted by α_1, using the Upwind scheme. The numerical results are generated using initial condition regularities $r_1 = 0, \ldots, 7$ and $r_1 \gg 1$ with $r_2 = 6$, by considering $N_x = 3^{\gamma}$ for $\gamma = 1, \ldots, 7$ and fixed $L = 4$ and calculating them through α_1 as in Equation (C.2). The values of γ listed in the table are the smaller of the two values of γ used to generate the N_x required to calculate α_1. These values were generated using $\mu_1 = \mu_2 = 1$ and $h_1 = h_2 = \frac{1}{2}$. Equation (5.120) was used to generate the order of convergence when $r_1 \gg 1$ and $r_2 = 6$.*

Upwind Scheme, $r_2 = 7$: $\alpha_1 = \log\left(\frac{E_{1,3N_x}}{E_{1,N_x}}\right)/\log(3)$									
γ \ r_1	0	1	2	3	4	5	6	7	$r_1 \gg 1$
1	6.2016×10^1	5.9984×10^1	5.8157×10^1	5.6542×10^1	5.5098×10^1	5.3783×10^1	5.2626×10^1	5.1776×10^1	5.1145×10^1
2	1.4709	5.7935×10^{-2}	−1.0105	−1.6029	−1.8394	−1.9215	−1.9502	−1.9588	−1.9588
3	1.1860	-4.6870×10^{-1}	−1.5864	−1.9364	−1.9886	−1.9959	−1.9974	−1.9977	−1.9977
4	1.0690	-7.4865×10^{-1}	−1.8450	−1.9937	−1.9993	−1.9997	−1.9998	−1.9998	−1.9998
5	1.0241	-8.8644×10^{-1}	−1.9454	−1.9995	−1.9999	−2.0000	−2.0000	−2.0000	−2.0000
6	1.0082	-9.5073×10^{-1}	−1.9912	−2.0207	−2.0148	−2.0076	−2.0034	−2.0018	−2.0020

Table C.8: *The numerical orders of convergence to zero with respect to N_x when $N_x = N_y$ for R_1, denoted by α_1, using the Upwind scheme. The numerical results are generated using initial condition regularities $r_1 = 0, \ldots, 7$ and $r_1 \gg 1$ with $r_2 = 7$, by considering $N_x = 3^\gamma$ for $\gamma = 1, \ldots, 7$ and fixed $L = 4$ and calculating them through α_1 as in Equation (C.2). The values of γ listed in the table are the smaller of the two values of γ used to generate the N_x required to calculate α_1. These values were generated using $\mu_1 = \mu_2 = 1$ and $h_1 = h_2 = \frac{1}{2}$. Equation (5.120) was used to generate the order of convergence when $r_1 \gg 1$ and $r_2 = 7$.*

Upwind Scheme, $r_2 \gg 1$: $\alpha_1 = \log\left(\frac{E_{1,3N_x}}{E_{1,N_x}}\right) / \log(3)$									
γ \ r_1	0	1	2	3	4	5	6	7	$r_1 \gg 1$
1	6.2016×10^{1}	5.9983×10^{1}	5.8156×10^{1}	5.6539×10^{1}	5.5086×10^{1}	5.3731×10^{1}	5.2427×10^{1}	5.1145×10^{1}	-5.4656×10^{-1}
2	1.4709	5.7980×10^{-2}	-1.0103	-1.6021	-1.8380	-1.9194	-1.9480	-1.9588	-1.9662
3	1.1860	-4.6870×10^{-1}	-1.5863	-1.9363	-1.9885	-1.9958	-1.9973	-1.9977	-1.9980
4	1.0690	-7.4865×10^{-1}	-1.8450	-1.9937	-1.9993	-1.9997	-1.9998	-1.9998	-1.9998
5	1.0241	-8.8644×10^{-1}	-1.9454	-1.9995	-2.0000	-2.0000	-2.0000	-2.0000	-2.0000
6	1.0082	-9.5073×10^{-1}	-1.9912	-2.0208	-2.0150	-2.0080	-2.0038	-2.0017	-2.0000

Table C.9: *The numerical orders of convergence to zero with respect to N_x when $N_x = N_y$ for R_1, denoted by α_1, using the Upwind scheme. The numerical results are generated using initial condition regularities $r_1 = 0, \ldots, 7$ and $r_1 \gg 1$ with $r_2 \gg 1$, by considering $N_x = 3^{\gamma}$ for $\gamma = 1, \ldots, 7$ and fixed $L = 4$ and calculating them through α_1 as in Equation (C.2). The values of γ listed in the table are the smaller of the two values of γ used to generate the N_x required to calculate α_1. These values were generated using $\mu_1 = \mu_2 = 1$ and $h_1 = h_2 = \frac{1}{2}$. Equations (5.121) and (5.122) were used to generate the order of convergence when $r_1 = 0, \ldots, 7$ and $r_2 \gg 1$ and $r_1 \gg 1$ and $r_2 \gg 1$ respectively.*

C.1.2 analytically for the Upwind scheme

The analytical order of convergence of R_1 for the Upwind scheme is calculated in the following.

$$E_1 = |R_1 - 0| = N_x N_y \sum_{p=2}^{\frac{N_x+1}{2}} \sum_{q=2}^{\frac{N_y+1}{2}} \frac{|1-\nu_{p,q}|^2}{|p-1|^{2(r_1+1)}|q-1|^{2(r_2+1)}}, \tag{C.3}$$

When $p = q$, $|1 - \nu_{p,q}| = 0$. When $p \neq q$ we use the following asymptotic approximation,

$$\begin{aligned} |1-\nu_{p,q}|^2 &= \left\{ K_{1a} L \left(\frac{p-1}{N_x}\right)^2 + K_{2a} L \left(\frac{p-1}{N_x}\right)\left(\frac{q-1}{N_y}\right) + K_{3a} L \left(\frac{q-1}{N_y}\right)^2 \right\}^2, \\ &= K_{1a}^2 L^2 \left(\frac{p-1}{N_x}\right)^4 + \left(K_{2a}^2 + 2K_{1a}K_{3a}\right) L^2 \left(\frac{p-1}{N_x}\right)^2 \left(\frac{q-1}{N_y}\right)^2 \\ &\quad + K_{3a}^2 L^2 \left(\frac{q-1}{N_y}\right)^4 + 2K_{1a}K_{2a} L^2 \left(\frac{p-1}{N_x}\right)^3 \left(\frac{q-1}{N_y}\right) \\ &\quad + 2K_{2a}K_{3a} L^2 \left(\frac{p-1}{N_x}\right)\left(\frac{q-1}{N_y}\right)^3. \end{aligned} \tag{C.4}$$

In order to calculate an approximation to R_1 for the Upwind scheme, we require the following bounds.

$$\sum_{p=1}^{\frac{N_x-1}{2}} \sum_{q=2}^{\frac{N_y+1}{2}} p^{2-2r_1} q^{-2-2r_2} < \begin{cases} \frac{\zeta(2r_2+2)}{24} N_x^3, & \text{for } r_1 = 0 \text{ and } r_2 \in \mathbb{N}_0, \\ \frac{\zeta(2r_2+2)}{2} N_x, & \text{for } r_1 = 1 \text{ and } r_2 \in \mathbb{N}_0, \\ \frac{\zeta(2r_1-2)\zeta(2r_2+2)}{24}, & \text{for } r_1 \geq 2 \text{ and } r_2 \in \mathbb{N}_0. \end{cases} \tag{C.5}$$

$$\sum_{p=1}^{\frac{N_x-1}{2}} \sum_{q=2}^{\frac{N_y+1}{2}} p^{-2r_1} q^{-2r_2} < \begin{cases} \frac{1}{4} N_x N_y, & \text{for } r_1 = 0 \text{ and } r_2 = 0, \\ \frac{\zeta(2r_1)}{2} N_y, & \text{for } r_1 \geq 1 \text{ and } r_2 = 0, \\ \frac{\zeta(2r_2)}{2} N_x, & \text{for } r_1 = 0 \text{ and } r_2 \geq 1, \\ \zeta(2r_1)\zeta(2r_2), & \text{for } r_1 \geq 1 \text{ and } r_2 \geq 1. \end{cases} \tag{C.6}$$

$$\sum_{p=1}^{\frac{N_x-1}{2}} \sum_{q=2}^{\frac{N_y+1}{2}} p^{-2-2r_1} q^{2-2r_2} < \begin{cases} \frac{\zeta(2r_1+2)}{24} N_y^3, & \text{for } r_1 \in \mathbb{N}_0 \text{ and } r_2 = 0, \\ \frac{\zeta(2r_1+2)}{2} N_y, & \text{for } r_1 \in \mathbb{N}_0 \text{ and } r_2 = 1, \\ \frac{\zeta(2r_1+2)\zeta(2r_2-2)}{24}, & \text{for } r_1 \in \mathbb{N}_0 \text{ and } r_2 \geq 2. \end{cases} \tag{C.7}$$

$$\sum_{p=1}^{\frac{N_x-1}{2}}\sum_{q=2}^{\frac{N_y+1}{2}} p^{1-2r_1}q^{-1-2r_2} < \begin{cases} \frac{2+\gamma}{2}N_x\log(N_y), & \text{for } r_1=0 \text{ and } r_2=0,\\ (2+\gamma)^2\log(N_x)\log(N_y), & \text{for } r_1=1 \text{ and } r_2=0,\\ \zeta(2r_1-1)(2+\gamma)\log(N_y), & \text{for } r_1\geq 2 \text{ and } r_2=0,\\ \frac{\zeta(2r_2+1)}{2}N_x, & \text{for } r_1=0 \text{ and } r_2\geq 1,\\ (2+\gamma)\zeta(2r_2+1)\log(N_x), & \text{for } r_1=1 \text{ and } r_2\geq 1,\\ \zeta(2r_1-1)\zeta(2r_2+1), & \text{for } r_1\geq 2 \text{ and } r_2\geq 1. \end{cases} \tag{C.8}$$

$$\sum_{p=1}^{\frac{N_x-1}{2}}\sum_{q=2}^{\frac{N_y+1}{2}} p^{-1-2r_1}q^{1-2r_2} < \begin{cases} \frac{2+\gamma}{2}\log(N_x)N_y, & \text{for } r_1=0 \text{ and } r_2=0,\\ \frac{\zeta(2r_1+1)}{2}N_y, & \text{for } r_1\geq 1 \text{ and } r_2=0,\\ (2+\gamma)^2\log(N_x)\log(N_y), & \text{for } r_1=0 \text{ and } r_2=1,\\ \zeta(2r_1+1)(2+\gamma)\log(N_y), & \text{for } r_1\geq 1 \text{ and } r_2=1,\\ (2+\gamma)\zeta(2r_2-1)\log(N_x), & \text{for } r_1=0 \text{ and } r_2\geq 2,\\ \zeta(2r_1+1)\zeta(2r_2-1), & \text{for } r_1\geq 1 \text{ and } r_2\geq 2. \end{cases} \tag{C.9}$$

We calculate an approximation for R_1 when $N_x = N_y$ by substituting the approximation in Equation (C.4) into Equation (C.3) and subtracting,

$$N_xN_y\sum_{p=2}^{\frac{N_x+1}{2}}\frac{|1-\nu_{p,p}|^2}{|p-1|^{2(r_1+r_2+2)}}, \tag{C.10}$$

with Equation (C.4) substituted in for $|1-\nu_{p,p}|^2$. The following bound is then required for these calculations,

$$\sum_{p=1}^{\frac{N_x-1}{2}} p^{-2(r_1+r_2+1)} < \begin{cases} \frac{N_x}{2}, & \text{for } r_1=0 \text{ and } r_2=0,\\ \zeta(2(r_1+r_2)), & \text{for } r_1 \text{ and } r_2 \text{ not both zero.} \end{cases} \tag{C.11}$$

When $N_x = N_y$ this results in,

$$\begin{aligned} &4A_4^2N_xN_y\sum_{p=1}^{\frac{N_x+1}{2}}\sum_{q=2}^{\frac{N_y+1}{2}}\frac{|1-\nu_{p,q}|^2}{|p-1|^{2(r_1+1)}|q-1|^{2(r_2+1)}}\\ =&\begin{cases} \mathcal{O}\left(L^2N_x\right), & \text{for } \min(r_1,r_2)=0,\\ \mathcal{O}\left(L^2N_x^{-1}\right), & \text{for } \min(r_1,r_2)=1,\\ \mathcal{O}\left(L^2N_x^{-2}\right), & \text{for } \min(r_1,r_2)\geq 2. \end{cases} \end{aligned} \tag{C.12}$$

C.1.3 with respect to N_xN_y for the Crank-Nicolson scheme

Crank-Nicolson scheme, $r_2 = 0$: $\alpha_1 = \log\left(\frac{E_{1,3N_x}}{E_{1,N_x}}\right)/\log(3)$									
γ \ r_1	0	1	2	3	4	5	6	7	$r_1 \gg 1$
1	1.6821	1.0081	8.4904×10^{-1}	8.1612×10^{-1}	8.0883×10^{-1}	8.0711×10^{-1}	8.0669×10^{-1}	8.0659×10^{-1}	8.0655×10^{-1}
2	1.0567	7.6511×10^{-1}	8.4236×10^{-1}	8.6178×10^{-1}	8.6604×10^{-1}	8.6703×10^{-1}	8.6727×10^{-1}	8.6733×10^{-1}	8.6735×10^{-1}
3	1.0018	9.5922×10^{-1}	9.8134×10^{-1}	9.8274×10^{-1}	9.8295×10^{-1}	9.8300×10^{-1}	9.8301×10^{-1}	9.8301×10^{-1}	9.8301×10^{-1}
4	9.9716×10^{-1}	9.9260×10^{-1}	9.9542×10^{-1}	9.9552×10^{-1}	9.9554×10^{-1}	9.9555×10^{-1}	9.9555×10^{-1}	9.9555×10^{-1}	9.9555×10^{-1}
5	9.9825×10^{-1}	9.9831×10^{-1}	9.9869×10^{-1}	9.9871×10^{-1}	9.9871×10^{-1}	9.9871×10^{-1}	9.9871×10^{-1}	9.9871×10^{-1}	9.9871×10^{-1}
6	9.9914×10^{-1}	9.9951×10^{-1}	9.9957×10^{-1}	9.9958×10^{-1}	9.9958×10^{-1}	9.9958×10^{-1}	9.9958×10^{-1}	9.9958×10^{-1}	9.9958×10^{-1}

Table C.10: *The numerical orders of convergence to zero with respect to N_x when $N_x = N_y$ for R_1, denoted by α_1, using the Crank-Nicolson scheme. The numerical results are generated using initial condition regularities $r_1 = 0, \ldots, 7$ and $r_1 \gg 1$ with $r_2 = 0$, by considering $N_x = 3^\gamma$ for $\gamma = 1, \ldots, 7$ and fixed $L = 4$ and calculating them through α_1 as in Equation (C.2). The values of γ listed in the table are the smaller of the two values of γ used to generate the N_x required to calculate α_1. These values were generated using $\mu_1 = \mu_2 = 1$ and $h_1 = h_2 = \frac{1}{2}$. Equation (5.123) was used to generate the order of convergence when $r_1 \gg 1$ and $r_2 = 0$.*

Crank-Nicolson Scheme, $r_2 = 1$: $\alpha_1 = \log\left(\frac{E_{1,3N_x}}{E_{1,N_x}}\right) / \log(3)$									
γ \ r_1	0	1	2	3	4	5	6	7	$r_1 \gg 1$
1	1.0081	-1.3610×10^{-1}	-5.1162×10^{-1}	-5.9546×10^{-1}	-6.1376×10^{-1}	-6.1800×10^{-1}	-6.1902×10^{-1}	-6.1927×10^{-1}	-6.1935×10^{-1}
2	7.6511×10^{-1}	-1.2401	-1.5167	-1.4751	-1.4625	-1.4595	-1.4588	-1.4586	-1.4586
3	9.5922×10^{-1}	-1.0682	-1.1015	-1.0755	-1.0727	-1.0722	-1.0721	-1.0721	-1.0721
4	9.9260×10^{-1}	-1.0152	-1.0171	-1.0135	-1.0133	-1.0132	-1.0132	-1.0132	-1.0132
5	9.9831×10^{-1}	-1.0041	-1.0040	-1.0036	-1.0036	-1.0036	-1.0036	-1.0036	-1.0036
6	9.9951×10^{-1}	-1.0013	-1.0012	-1.0011	-1.0011	-1.0011	-1.0011	-1.0011	-1.0011

Table C.11: *The numerical orders of convergence to zero with respect to N_x when $N_x = N_y$ for R_1, denoted by α_1, using the Crank-Nicolson scheme. The numerical results are generated using initial condition regularities $r_1 = 0, \ldots, 7$ and $r_1 \gg 1$ with $r_2 = 1$, by considering $N_x = 3^\gamma$ for $\gamma = 1, \ldots, 7$ and fixed $L = 4$ and calculating them through α_1 as in Equation (C.2). The values of γ listed in the table are the smaller of the two values of γ used to generate the N_x required to calculate α_1. These values were generated using $\mu_1 = \mu_2 = 1$ and $h_1 = h_2 = \frac{1}{2}$. Equation (5.123) was used to generate the order of convergence when $r_1 \gg 1$ and $r_2 = 1$.*

Crank-Nicolson Scheme, $r_2 = 2$: $\alpha_1 = \log\left(\frac{E_{1,3N_x}}{E_{1,N_x}}\right)/\log(3)$									
γ \ r_1	0	1	2	3	4	5	6	7	$r_1 \gg 1$
1	8.4903×10^{-1}	-5.1162×10^{-1}	-1.0648	-1.2031	-1.2341	-1.2412	-1.2429	-1.2433	-1.2435
2	8.4236×10^{-1}	-1.5167	-3.1455	-3.3550	-3.3635	-3.3634	-3.3633	-3.3633	-3.3632
3	9.8134×10^{-1}	-1.1015	-3.1821	-3.3389	-3.3244	-3.3213	-3.3206	-3.3204	-3.3204
4	9.9542×10^{-1}	-1.0171	-3.0970	-3.1809	-3.1684	-3.1664	-3.1660	-3.1659	-3.1659
5	9.9869×10^{-1}	-1.0040	-3.0421	-3.0770	-3.0710	-3.0701	-3.0700	-3.0699	-3.0699
6	9.9957×10^{-1}	-1.0012	-3.0168	-3.0295	-3.0272	-3.0269	-3.0268	-3.0268	-3.0268

Table C.12: *The numerical orders of convergence to zero with respect to N_x when $N_x = N_y$ for R_1, denoted by α_1, using the Crank-Nicolson scheme. The numerical results are generated using initial condition regularities $r_1 = 0, \ldots, 7$ and $r_1 \gg 1$ with $r_2 = 2$, by considering $N_x = 3^\gamma$ for $\gamma = 1, \ldots, 7$ and fixed $L = 4$ and calculating them through α_1 as in Equation (C.2). The values of γ listed in the table are the smaller of the two values of γ used to generate the N_x required to calculate α_1. These values were generated using $\mu_1 = \mu_2 = 1$ and $h_1 = h_2 = \frac{1}{2}$. Equation (5.123) was used to generate the order of convergence when $r_1 \gg 1$ and $r_2 = 2$.*

Crank-Nicolson Scheme, $r_2 = 3$: $\alpha_1 = \log\left(\frac{E_{1,3N_x}}{E_{1,N_x}}\right)/\log(3)$									
γ \ r_1	0	1	2	3	4	5	6	7	$r_1 \gg 1$
1	8.1612×10^{-1}	-5.9546×10^{-1}	-1.2031	-1.3613	-1.3970	-1.4052	-1.4072	-1.4077	-1.4079
2	8.6178×10^{-1}	-1.4751	-3.3550	-3.7391	-3.7712	-3.7752	-3.7759	-3.7761	-3.7761
3	9.8274×10^{-1}	-1.0755	-3.3389	-3.9499	-3.9621	-3.9627	-3.9627	-3.9627	-3.9627
4	9.9552×10^{-1}	-1.0135	-3.1809	-3.9883	-3.9924	-3.9924	-3.9924	-3.9924	-3.9924
5	9.9871×10^{-1}	-1.0036	-3.0770	-3.9967	-3.9981	-3.9981	-3.9981	-3.9981	-3.9981
6	9.9958×10^{-1}	-1.0011	-3.0295	-3.9990	-3.9994	-3.9994	-3.9994	-3.9994	-3.9994

Table C.13: *The numerical orders of convergence to zero with respect to N_x when $N_x = N_y$ for R_1, denoted by α_1, using the Crank-Nicolson scheme. The numerical results are generated using initial condition regularities $r_1 = 0, \ldots, 7$ and $r_1 \gg 1$ with $r_2 = 3$, by considering $N_x = 3^\gamma$ for $\gamma = 1, \ldots, 7$ and fixed $L = 4$ and calculating them through α_1 as in Equation (C.2). The values of γ listed in the table are the smaller of the two values of γ used to generate the N_x required to calculate α_1. These values were generated using $\mu_1 = \mu_2 = 1$ and $h_1 = h_2 = \frac{1}{2}$. Equation (5.123) was used to generate the order of convergence when $r_1 \gg 1$ and $r_2 = 3$.*

Crank-Nicolson Scheme, $r_2 = 4$: $\alpha_1 = \log\left(\frac{E_{1,3N_x}}{E_{1,N_x}}\right)/\log(3)$									
γ \ r_1	0	1	2	3	4	5	6	7	$r_1 \gg 1$
1	8.0883×10^{-1}	-6.1376×10^{-1}	-1.2341	-1.3970	-1.4338	-1.4424	-1.4444	-1.4449	-1.4451
2	8.6604×10^{-1}	-1.4625	-3.3635	-3.7712	-3.8067	-3.8113	-3.8122	-3.8124	-3.8124
3	9.8295×10^{-1}	-1.0727	-3.3244	-3.9621	-3.9765	-3.9773	-3.9774	-3.9774	-3.9774
4	9.9554×10^{-1}	-1.0133	-3.1684	-3.9924	-3.9973	-3.9974	-3.9974	-3.9974	-3.9975
5	9.9871×10^{-1}	-1.0036	-3.0710	-3.9981	-3.9997	-3.9997	-3.9997	-3.9997	-3.9997
6	9.9958×10^{-1}	-1.0011	-3.0272	-3.9994	-4.0000	-4.0000	-4.0000	-4.0000	-4.0000

Table C.14: *The numerical orders of convergence to zero with respect to N_x when $N_x = N_y$ for R_1, denoted by α_1, using the Crank-Nicolson scheme. The numerical results are generated using initial condition regularities $r_1 = 0, \ldots, 7$ and $r_1 \gg 1$ with $r_2 = 4$, by considering $N_x = 3^\gamma$ for $\gamma = 1, \ldots, 7$ and fixed $L = 4$ and calculating them through α_1 as in Equation (C.2). The values of γ listed in the table are the smaller of the two values of γ used to generate the N_x required to calculate α_1. These values were generated using $\mu_1 = \mu_2 = 1$ and $h_1 = h_2 = \frac{1}{2}$. Equation (5.123) was used to generate the order of convergence when $r_1 \gg 1$ and $r_2 = 4$.*

Crank-Nicolson Scheme, $r_2 = 5$: $\alpha_1 = \log\left(\frac{E_{1,3N_x}}{E_{1,N_x}}\right)/\log(3)$									
γ \ r_1	0	1	2	3	4	5	6	7	$r_1 \gg 1$
1	8.0711×10^{-1}	-6.1800×10^{-1}	-1.2412	-1.4052	-1.4424	-1.4510	-1.4530	-1.4535	-1.4537
2	8.6703×10^{-1}	-1.4595	-3.3634	-3.7752	-3.8113	-3.8160	-3.8168	-3.8170	-3.8171
3	9.8300×10^{-1}	-1.0722	-3.3213	-3.9627	-3.9773	-3.9781	-3.9782	-3.9782	-3.9782
4	9.9555×10^{-1}	-1.0132	-3.1664	-3.9924	-3.9974	-3.9975	-3.9976	-3.9976	-3.9976
5	9.9871×10^{-1}	-1.0036	-3.0701	-3.9981	-3.9997	-3.9997	-3.9997	-3.9997	-3.9997
6	9.9958×10^{-1}	-1.0011	-3.0269	-3.9994	-4.0000	-4.0000	-4.0000	-4.0000	-4.0000

Table C.15: *The numerical orders of convergence to zero with respect to N_x when $N_x = N_y$ for R_1, denoted by α_1, using the Crank-Nicolson scheme. The numerical results are generated using initial condition regularities $r_1 = 0, \ldots, 7$ and $r_1 \gg 1$ with $r_2 = 5$, by considering $N_x = 3^\gamma$ for $\gamma = 1, \ldots, 7$ and fixed $L = 4$ and calculating them through α_1 as in Equation (C.2). The values of γ listed in the table are the smaller of the two values of γ used to generate the N_x required to calculate α_1. These values were generated using $\mu_1 = \mu_2 = 1$ and $h_1 = h_2 = \frac{1}{2}$. Equation (5.123) was used to generate the order of convergence when $r_1 \gg 1$ and $r_2 = 5$.*

Crank-Nicolson Scheme, $r_2 = 6$: $\alpha_1 = \log\left(\frac{E_{1,3N_x}}{E_{1,N_x}}\right)/\log(3)$									
γ \ r_1	0	1	2	3	4	5	6	7	$r_1 \gg 1$
1	8.0669×10^{-1}	-6.1902×10^{-1}	-1.2429	-1.4072	-1.4444	-1.4530	-1.4551	-1.4556	-1.4558
2	8.6727×10^{-1}	-1.4588	-3.3633	-3.7759	-3.8122	-3.8168	-3.8177	-3.8179	-3.8180
3	9.8301×10^{-1}	-1.0721	-3.3206	-3.9627	-3.9774	-3.9782	-3.9783	-3.9783	-3.9783
4	9.9555×10^{-1}	-1.0132	-3.1660	-3.9924	-3.9974	-3.9976	-3.9976	-3.9976	-3.9976
5	9.9871×10^{-1}	-1.0036	-3.0700	-3.9981	-3.9997	-3.9997	-3.9997	-3.9997	-3.9997
6	9.9958×10^{-1}	-1.0011	-3.0268	-3.9994	-4.0000	-4.0000	-4.0000	-4.0000	-4.0000

Table C.16: *The numerical orders of convergence to zero with respect to N_x when $N_x = N_y$ for R_1, denoted by α_1, using the Crank-Nicolson scheme. The numerical results are generated using initial condition regularities $r_1 = 0, \ldots, 7$ and $r_1 \gg 1$ with $r_2 = 6$, by considering $N_x = 3^\gamma$ for $\gamma = 1, \ldots, 7$ and fixed $L = 4$ and calculating them through α_1 as in Equation (C.2). The values of γ listed in the table are the smaller of the two values of γ used to generate the N_x required to calculate α_1. These values were generated using $\mu_1 = \mu_2 = 1$ and $h_1 = h_2 = \frac{1}{2}$. Equation (5.123) was used to generate the order of convergence when $r_1 \gg 1$ and $r_2 = 6$.*

Crank-Nicolson Scheme, $r_2 = 7$: $\alpha_1 = \log\left(\frac{E_{1,3N_x}}{E_{1,N_x}}\right)/\log(3)$									
γ \ r_1	0	1	2	3	4	5	6	7	$r_1 \gg 1$
1	8.0659×10^{-1}	-6.1927×10^{-1}	-1.2433	-1.4077	-1.4449	-1.4536	-1.4556	-1.4561	-1.4563
2	8.6733×10^{-1}	-1.4586	-3.3633	-3.7761	-3.8124	-3.8170	-3.8179	-3.8181	-3.8182
3	9.8301×10^{-1}	-1.0721	-3.3204	-3.9627	-3.9774	-3.9782	-3.9783	-3.9784	-3.9784
4	9.9555×10^{-1}	-1.0132	-3.1659	-3.9924	-3.9974	-3.9976	-3.9976	-3.9976	-3.9976
5	9.9871×10^{-1}	-1.0036	-3.0699	-3.9981	-3.9997	-3.9997	-3.9997	-3.9997	-3.9997
6	9.9958×10^{-1}	-1.0011	-3.0268	-3.9994	-4.0000	-4.0000	-4.0000	-4.0000	-4.0000

Table C.17: *The numerical orders of convergence to zero with respect to N_x when $N_x = N_y$ for R_1, denoted by α_1, using the Crank-Nicolson scheme. The numerical results are generated using initial condition regularities $r_1 = 0, \ldots, 7$ and $r_1 \gg 1$ with $r_2 = 7$, by considering $N_x = 3^\gamma$ for $\gamma = 1, \ldots, 7$ and fixed $L = 4$ and calculating them through α_1 as in Equation (C.2). The values of γ listed in the table are the smaller of the two values of γ used to generate the N_x required to calculate α_1. These values were generated using $\mu_1 = \mu_2 = 1$ and $h_1 = h_2 = \frac{1}{2}$. Equation (5.123) was used to generate the order of convergence when $r_1 \gg 1$ and $r_2 = 7$.*

Crank-Nicolson Scheme, $r_2 \gg 1$: $\alpha_1 = \log\left(\frac{E_{1,3N_x}}{E_{1,N_x}}\right) / \log(3)$									
γ \ r_1	0	1	2	3	4	5	6	7	$r_1 \gg 1$
1	8.0655×10^{-1}	-6.1935×10^{-1}	-1.2435	-1.4079	-1.4451	-1.4537	-1.4558	-1.4563	-1.4565
2	8.6735×10^{-1}	-1.4586	-3.3632	-3.7761	-3.8124	-3.8171	-3.8180	-3.8182	-3.8182
3	9.8301×10^{-1}	-1.0721	-3.3204	-3.9627	-3.9774	-3.9782	-3.9783	-3.9784	-3.9784
4	9.9555×10^{-1}	-1.0132	-3.1659	-3.9924	-3.9975	-3.9976	-3.9976	-3.9976	-3.9976
5	9.9871×10^{-1}	-1.0036	-3.0699	-3.9981	-3.9997	-3.9997	-3.9997	-3.9997	-3.9997
6	9.9958×10^{-1}	-1.0011	-3.0268	-3.9994	-4.0000	-4.0000	-4.0000	-4.0000	-4.0000

Table C.18: *The numerical orders of convergence to zero with respect to N_x when $N_x = N_y$ for R_1, denoted by α_1, using the Crank-Nicolson scheme. The numerical results are generated using initial condition regularities $r_1 = 0, \ldots, 7$ and $r_1 \gg 1$ with $r_2 \gg 1$, by considering $N_x = 3^\gamma$ for $\gamma = 1, \ldots, 7$ and fixed $L = 4$ and calculating them through α_1 as in Equation (C.2). The values of γ listed in the table are the smaller of the two values of γ used to generate the N_x required to calculate α_1. These values were generated using $\mu_1 = \mu_2 = 1$ and $h_1 = h_2 = \frac{1}{2}$. Equations (5.124) and (5.125) were used to generate the order of convergence when $r_1 = 0, \ldots, 7$ and $r_2 \gg 1$ and $r_1 \gg 1$ and $r_2 \gg 1$ respectively.*

C.1.4 analytically for the Crank-Nicolson scheme

The analytical order of convergence of R_1 for the Crank-Nicolson scheme is calculated in the following.

$$E_1 = |R_1 - 0| = N_x N_y \sum_{p=2}^{\frac{N_x+1}{2}} \sum_{q=2}^{\frac{N_y+1}{2}} \frac{|1-\nu_{p,q}|^2}{|p-1|^{2(r_1+1)}|q-1|^{2(r_2+1)}}. \tag{C.13}$$

We use the following asymptotic approximation,

$$\begin{aligned}
& |1-\nu_{p,q}|^2 \\
= & \left\{ K_{1b} L \left(\frac{p-1}{N_x}\right)^3 + K_{2b} L \left(\frac{p-1}{N_x}\right)^2 \left(\frac{q-1}{N_y}\right) + K_{3b} L \left(\frac{p-1}{N_x}\right) \left(\frac{q-1}{N_y}\right)^2 \right. \\
& \left. + K_{4b} L \left(\frac{q-1}{N_y}\right)^3 \right\}^2, \\
= & K_{1b}^2 L^2 \left(\frac{p-1}{N_x}\right)^6 + 2K_{1b}K_{2b}L^2 \left(\frac{p-1}{N_x}\right)^5 \left(\frac{q-1}{N_y}\right) \\
& + \left(K_{2b}^2 + 2K_{1b}K_{3b}\right) L^2 \left(\frac{p-1}{N_x}\right)^4 \left(\frac{q-1}{N_y}\right)^2 \\
& + 2\left(K_{1b}K_{4b} + K_{2b}K_{3b}\right) L^2 \left(\frac{p-1}{N_x}\right)^3 \left(\frac{q-1}{N_y}\right)^3 \\
& + \left(K_{3b}^2 + 2K_{2b}K_{4b}\right) L^2 \left(\frac{p-1}{N_x}\right)^2 \left(\frac{q-1}{N_y}\right)^4 + 2K_{3b}K_{4b}L^2 \left(\frac{p-1}{N_x}\right) \left(\frac{q-1}{N_y}\right)^5 \\
& + K_{4b}^2 L^2 \left(\frac{q-1}{N_y}\right)^6.
\end{aligned} \tag{C.14}$$

In order to calculate an approximation to R_1 for the Crank-Nicolson scheme, we require the following bounds.

$$\sum_{p=1}^{\frac{N_x-1}{2}} \sum_{q=2}^{\frac{N_y+1}{2}} p^{4-2r_1} q^{-2-2r_2} < \begin{cases} \frac{\zeta(2r_2+2)}{160} N_x^5, & \text{for } r_1 = 0 \text{ and } r_2 \in \mathbb{N}_0, \\ \frac{\zeta(2r_2+2)}{24} N_x^3, & \text{for } r_1 = 1 \text{ and } r_2 \in \mathbb{N}_0, \\ \frac{\zeta(2r_2+2)}{2} N_x, & \text{for } r_1 = 2 \text{ and } r_2 \in \mathbb{N}_0, \\ \zeta(2r_1-4)\zeta(2r_2+2), & \text{for } r_1 \geq 3 \text{ and } r_2 \in \mathbb{N}_0. \end{cases} \tag{C.15}$$

$$\sum_{p=1}^{\frac{N_x-1}{2}}\sum_{q=2}^{\frac{N_y+1}{2}} p^{3-2r_1}q^{-1-2r_2} < \begin{cases} \frac{3(2+\gamma)}{8}N_x^4\log(N_y), & \text{for } r_1=0 \text{ and } r_2=0,\\ \frac{3(2+\gamma)}{4}N_x^2\log(N_y), & \text{for } r_1=1 \text{ and } r_2=0,\\ (2+\gamma)^2\log(N_x)\log(N_y), & \text{for } r_1=2 \text{ and } r_2=0,\\ (2+\gamma)\zeta(2r_1-3)\log(N_y), & \text{for } r_1\geq 3 \text{ and } r_2=0,\\ \frac{3\zeta(2r_2+1)}{8}N_x^4, & \text{for } r_1=0 \text{ and } r_2\geq 1,\\ \frac{3\zeta(2r_2+1)}{4}N_x^2, & \text{for } r_1=1 \text{ and } r_2\geq 1,\\ (2+\gamma)\zeta(2r_2+1)\log(N_x), & \text{for } r_1=2 \text{ and } r_2\geq 1,\\ \zeta(2r_1-3)\zeta(2r_2+1), & \text{for } r_1\geq 3 \text{ and } r_2\geq 1. \end{cases} \tag{C.16}$$

$$\sum_{p=1}^{\frac{N_x-1}{2}}\sum_{q=2}^{\frac{N_y+1}{2}} p^{2-2r_1}q^{-2r_2} < \begin{cases} \frac{1}{48}N_x^3N_y, & \text{for } r_1=0 \text{ and } r_2=0,\\ \frac{1}{4}N_xN_y, & \text{for } r_1=1 \text{ and } r_2=0,\\ \frac{\zeta(2r_1-2)}{2}N_y, & \text{for } r_1\geq 2 \text{ and } r_2=0,\\ \frac{\zeta(2r_2)}{24}N_x^3, & \text{for } r_1=0 \text{ and } r_2\geq 1,\\ \frac{\zeta(2r_2)}{2}N_x, & \text{for } r_1=1 \text{ and } r_2\geq 1,\\ \zeta(2r_1-2)\zeta(2r_2), & \text{for } r_1\geq 2 \text{ and } r_2\geq 1. \end{cases} \tag{C.17}$$

$$\sum_{p=1}^{\frac{N_x-1}{2}}\sum_{q=2}^{\frac{N_y+1}{2}} p^{1-2r_1}q^{1-2r_2} < \begin{cases} \frac{9}{16}N_x^2N_y^2, & \text{for } r_1=0 \text{ and } r_2=0,\\ \frac{3(2+\gamma)}{4}\log(N_x)N_y^2, & \text{for } r_1=1 \text{ and } r_2=0,\\ \frac{3\zeta(2r_1-1)}{4}N_y^2, & \text{for } r_1\geq 2 \text{ and } r_2=0,\\ \frac{3(2+\gamma)}{4}N_x^2\log(N_y), & \text{for } r_1=0 \text{ and } r_2\geq 1,\\ (2+\gamma)^2\log(N_x)\log(N_y), & \text{for } r_1=1 \text{ and } r_2\geq 1,\\ (2+\gamma)\zeta(2r_1-1)\log(N_y), & \text{for } r_1\geq 2 \text{ and } r_2\geq 1,\\ \frac{3\zeta(2r_2-1)}{4}N_x^2, & \text{for } r_1=0 \text{ and } r_2\geq 2,\\ (2+\gamma)\zeta(2r_2-1)\log(N_x), & \text{for } r_1=1 \text{ and } r_2\geq 2,\\ \zeta(2r_1-1)\zeta(2r_2-1), & \text{for } r_1\geq 2 \text{ and } r_2\geq 2. \end{cases} \tag{C.18}$$

$$\sum_{p=1}^{\frac{N_x-1}{2}}\sum_{q=2}^{\frac{N_y+1}{2}} p^{-2r_1}q^{2-2r_2} < \begin{cases} \frac{1}{48}N_xN_y^3, & \text{for } r_1=0 \text{ and } r_2=0,\\ \frac{\zeta(2r_1)}{24}N_y^3, & \text{for } r_1\geq 1 \text{ and } r_2=0,\\ \frac{1}{4}N_xN_y, & \text{for } r_1=0 \text{ and } r_2=1,\\ \frac{\zeta(2r_1)}{2}N_y, & \text{for } r_1\geq 1 \text{ and } r_2=1,\\ \frac{\zeta(2r_2-2)}{2}N_x, & \text{for } r_1=0 \text{ and } r_2\geq 2,\\ \zeta(2r_1)\zeta(2r_2-2), & \text{for } r_1\geq 1 \text{ and } r_2\geq 2. \end{cases} \tag{C.19}$$

$$\sum_{p=1}^{\frac{N_x-1}{2}}\sum_{q=2}^{\frac{N_y+1}{2}} p^{-1-2r_1}q^{3-2r_2} < \begin{cases} \frac{3(2+\gamma)}{8}\log(N_x)N_y^4, & \text{for } r_1=0 \text{ and } r_2=0,\\ \frac{3\zeta(2r_1+1)}{8}N_y^4, & \text{for } r_1=1 \text{ and } r_2=0,\\ \frac{3(2+\gamma)}{4}\log(N_x)N_y^2, & \text{for } r_1=0 \text{ and } r_2=1,\\ \frac{3\zeta(2r_1+1)}{4}N_y^2, & \text{for } r_1\geq 1 \text{ and } r_2=1,\\ (2+\gamma)^2\log(N_x)\log(N_y), & \text{for } r_1=0 \text{ and } r_2=2,\\ (2+\gamma)\zeta(2r_1+1)\log(N_y), & \text{for } r_1\geq 1 \text{ and } r_2=2,\\ (2+\gamma)\zeta(2r_2-3)\log(N_x), & \text{for } r_1=0 \text{ and } r_2\geq 3,\\ \zeta(2r_1+1)\zeta(2r_2-3), & \text{for } r_1\geq 1 \text{ and } r_2\geq 3. \end{cases} \tag{C.20}$$

$$\sum_{p=1}^{\frac{N_x-1}{2}}\sum_{q=2}^{\frac{N_y+1}{2}} p^{-2-2r_1}q^{4-2r_2} < \begin{cases} \frac{\zeta(2r_1+2)}{160}N_y^5, & \text{for } r_1\in\mathbb{N}_0 \text{ and } r_2=0,\\ \frac{\zeta(2r_1+2)}{24}N_y^3, & \text{for } r_1\in\mathbb{N}_0 \text{ and } r_2=1,\\ \frac{\zeta(2r_1+2)}{2}N_y, & \text{for } r_1\in\mathbb{N}_0 \text{ and } r_2=2,\\ \zeta(2r_1+2)\zeta(2r_2+4), & \text{for } r_1\in\mathbb{N}_0 \text{ and } r_2\geq 3. \end{cases} \tag{C.21}$$

We calculate an approximation for R_1 when $N_x = N_y$ by substituting the approximation in Equation (C.14) into Equation (C.13) which results in,

$$4A_4^2N_xN_y\sum_{p=2}^{\frac{N_x+1}{2}}\sum_{q=2}^{\frac{N_y+1}{2}}\frac{|1-\nu_{p,q}|^2}{|p-1|^{2(r_1+1)}|q-1|^{2(r_2+1)}} = \begin{cases} \mathcal{O}(L^2N_x), & \text{for } \min(r_1,r_2)=0,\\ \mathcal{O}(L^2N_x^{-1}), & \text{for } \min(r_1,r_2)=1,\\ \mathcal{O}(L^2N_x^{-3}), & \text{for } \min(r_1,r_2)=2,\\ \mathcal{O}(L^2N_x^{-4}), & \text{for } \min(r_1,r_2)\geq 3. \end{cases} \tag{C.22}$$

C.2 The orders of convergence for R_2...

C.2.1 with respect to N_xN_y for the Upwind scheme

Upwind Scheme: $\alpha_2 = \log\left(\frac{E_{2,3N_x}}{E_{2,N_x}}\right)/\log(3)$									
γ \ r_1	0	1	2	3	4	5	6	7	$r_1 \gg 1$
1	1.1202	1.4123×10^{-1}	-2.7189×10^{-1}	-3.9227×10^{-1}	-4.2274×10^{-1}	-4.3027×10^{-1}	-4.3213×10^{-1}	-4.3259×10^{-1}	-4.3275×10^{1}
2	1.0004	-8.4580×10^{-1}	-1.7700	-1.9371	-1.9608	-1.9651	-1.9660	-1.9662	-1.9662
3	1.0000	-9.5391×10^{-1}	-1.9335	-1.9957	-1.9978	-1.9979	-1.9980	-1.9980	-1.9980
4	1.0000	-9.8514×10^{-1}	-1.9788	-1.9996	-1.9998	-1.9998	-1.9998	-1.9998	-1.9998
5	1.0000	-9.9475×10^{-1}	-1.9529	-1.9396	-1.9358	-1.9350	-1.9348	-1.9348	-1.9347
6	1.0000	-9.0946×10^{-1}	1.2255	1.5740	1.6255	1.6364	1.6390	1.6396	1.6398

Table C.19: *The numerical orders of convergence to zero with respect to N_x when $N_x = N_y$ for R_2, denoted by α_2, using the Upwind scheme. The numerical results are generated using initial condition regularities $r_1 = 0, \ldots, 7$ and $r_1 \gg 1$, by considering $N_x = 3^\gamma$ for $\gamma = 1, \ldots, 7$ and fixed $L = 4$ and calculating them through α_2 as in Equation (C.2). The values of γ listed in the table are the smaller of the two values of γ used to generate the N_x required to calculate α_2. These values were generated using $\mu_1 = \mu_2 = 1$ and $h_1 = h_2 = \frac{1}{2}$. Equation (5.128) was used to generate the order of convergence when $r_1 \gg 1$.*

C.2.2 analytically for the Upwind scheme

The analytical order of convergence of R_2 for the Upwind scheme is calculated in the following.

$$\begin{aligned}
&\sum_{p=2}^{\frac{N_x-1}{2}} \frac{|1-\nu_{p,1}|^2}{|p-1|^{2(r_1+1)}}, \\
= \quad & \frac{K_{1a}^2 L^2}{N_x^4} \sum_{p=1}^{\frac{N_x-1}{2}} p^{2-2r_1}, \\
< \quad & \begin{cases} \frac{K_{1a}^2}{24} L^2 N_x^{-1}, & \text{for } r_1 = 0, \\ \frac{K_{1a}^2}{2} L^2 N_x^{-3}, & \text{for } r_1 = 1, \\ K_{1a}^2 \zeta(2r_1-2) L^2 N_x^{-4}, & \text{for } r_1 \geq 2, \end{cases}
\end{aligned}$$

$$\Rightarrow \quad E_2 = N_x N_y \sum_{p=2}^{\frac{N_x-1}{2}} \frac{|1-\nu_{p,1}|^2}{|p-1|^{2(r_1+1)}} < \begin{cases} \frac{K_{1a}^2}{24} L^2 N_y, & \text{for } r_1 = 0, \\ \frac{K_{1a}^2}{2} L^2 N_x^{-2} N_y, & \text{for } r_1 = 1, \\ \frac{K_{1a}^2 \zeta(2r_1-2)}{4} L^2 N_x^{-3} N_y, & \text{for } r_1 \geq 2, \end{cases} \tag{C.23}$$

$$\Rightarrow \quad 2N_x N_y A_3^2 \sum_{p=2}^{\frac{N_x-1}{2}} \frac{|1-\nu_{p,1}|^2}{|p-1|^{2(r_1+1)}} = \begin{cases} \mathcal{O}(L^2 N_y), & \text{for } r_1 = 0, \\ \mathcal{O}(L^2 N_x^{-2} N_y), & \text{for } r_1 = 1, \\ \mathcal{O}(L^2 N_x^{-3} N_y), & \text{for } r_1 \geq 2. \end{cases} \tag{C.24}$$

When $N_x = N_y$ this results in,

$$2N_x N_y A_3^2 \sum_{p=2}^{\frac{N_x-1}{2}} \frac{|1-\nu_{p,1}|^2}{|p-1|^{2(r_1+1)}} = \begin{cases} \mathcal{O}(L^2 N_x), & \text{for } r_1 = 0, \\ \mathcal{O}(L^2 N_x^{-1}), & \text{for } r_1 = 1, \\ \mathcal{O}(L^2 N_x^{-2}), & \text{for } r_1 \geq 2. \end{cases} \tag{C.25}$$

C.2.3 with respect to N_xN_y for the Crank-Nicolson scheme

Crank-Nicolson Scheme: $\alpha_2 = \log\left(\frac{E_{2,3N_x}}{E_{2,N_x}}\right)/\log(3)$									
γ \ r_1	0	1	2	3	4	5	6	7	$r_1 \gg 1$
1	6.3590×10^{-1}	-1.1610	-2.3905	-2.9133	-3.0578	-3.0925	-3.1008	-3.1029	-3.1035
2	1.0346	-9.8670×10^{-1}	-2.8564	-3.7582	-3.8994	-3.9174	-3.9206	-3.9212	-3.9214
3	1.0015	-9.9948×10^{-1}	-2.9574	-3.9261	-3.9873	-3.9904	-3.9908	-3.9909	-3.9909
4	9.9980×10^{-1}	-1.0001	-2.9863	-3.9760	-3.9985	-3.9989	-3.9990	-3.9990	-3.9990
5	1.0000	-1.0000	-2.9955	-3.9921	-3.9998	-3.9999	-3.9999	-3.9999	-3.9999
6	1.0000	-1.0000	-2.9985	-3.9974	-4.0000	-4.0000	-4.0000	-4.0000	-4.0000

Table C.20: *The numerical orders of convergence to zero with respect to N_x when $N_x = N_y$ for R_2, denoted by α_2, using the Crank-Nicolson scheme. The numerical results are generated using initial condition regularities $r_1 = 0, \ldots, 7$ and $r_1 \gg 1$, by considering $N_x = 3^\gamma$ for $\gamma = 1, \ldots, 7$ and fixed $L = 4$ and calculating them through α_2 as in Equation (C.2). The values of γ listed in the table are the smaller of the two values of γ used to generate the N_x required to calculate α_2. These values were generated using $\mu_1 = \mu_2 = 1$ and $h_1 = h_2 = \frac{1}{2}$. Equation (5.128) was used to generate the order of convergence when $r_1 \gg 1$.*

C.2.4 analytically for the Crank-Nicolson scheme

The analytical order of convergence of R_2 for the Upwind scheme is calculated in the following.

$$
\begin{aligned}
& \sum_{p=2}^{\frac{N_x-1}{2}} \frac{|1-\nu_{p,1}|^2}{|p-1|^{2(r_1+1)}}, \\
= \; & \frac{K_{1b}^2 L^2}{N_x^6} \sum_{p=1}^{\frac{N_x-1}{2}} p^{4-2r_1}, \\
< \; & \begin{cases} \frac{K_{1b}^2}{160} L^2 N_x^{-1}, & \text{for } r_1 = 0, \\ \frac{K_{1b}^2}{6} L^2 N_x^{-3}, & \text{for } r_1 = 1, \\ \frac{K_{1b}^2}{2} L^2 N_x^{-5}, & \text{for } r_1 = 2, \\ K_{1b}^2 \zeta(2r_1-4) L^2 N_x^{-6}, & \text{for } r_1 \geq 3, \end{cases}
\end{aligned}
$$

$$
\Rightarrow \quad E_2 = N_x N_y \sum_{p=2}^{\frac{N_x-1}{2}} \frac{|1-\nu_{p,1}|^2}{|p-1|^{2(r_1+1)}} < \begin{cases} \frac{K_{1b}^2}{160} L^2 N_y, & \text{for } r_1 = 0, \\ \frac{K_{1b}^2}{6} L^2 N_x^{-2} N_y, & \text{for } r_1 = 1, \\ \frac{K_{1b}^2}{2} L^2 N_x^{-4} N_y, & \text{for } r_1 = 2, \\ 2K_{1b}^2 \zeta(2r_1-4) L^2 N_x^{-5} N_y, & \text{for } r_1 \geq 3, \end{cases} \tag{C.26}
$$

$$
\Rightarrow \quad 2N_x N_y A_3^2 \sum_{p=2}^{\frac{N_x-1}{2}} \frac{|1-\nu_{p,1}|^2}{|p-1|^{2(r_1+1)}} = \begin{cases} \mathcal{O}(L^2 N_y), & \text{for } r_1 = 0, \\ \mathcal{O}(L^2 N_x^{-2} N_y), & \text{for } r_1 = 1, \\ \mathcal{O}(L^2 N_x^{-4} N_y), & \text{for } r_1 = 2, \\ \mathcal{O}(L^2 N_x^{-5} N_y), & \text{for } r_1 \geq 3. \end{cases} \tag{C.27}
$$

When $N_x = N_y$ this results in,

$$
2N_x N_y A_3^2 \sum_{p=2}^{\frac{N_x-1}{2}} \frac{|1-\nu_{p,1}|^2}{|p-1|^{2(r_1+1)}} = \begin{cases} \mathcal{O}(L^2 N_x), & \text{for } r_1 = 0, \\ \mathcal{O}(L^2 N_x^{-1}), & \text{for } r_1 = 1, \\ \mathcal{O}(L^2 N_x^{-3}), & \text{for } r_1 = 2, \\ \mathcal{O}(L^2 N_x^{-4}), & \text{for } r_1 \geq 3. \end{cases} \tag{C.28}
$$

C.3 The orders of convergence for R_3...

C.3.1 with respect to N_xN_y for the Upwind scheme

Upwind Scheme: $\alpha_3 = \log\left(\frac{E_{3,3N_x}}{E_{3,N_x}}\right)/\log(3)$									
γ \ r_2	0	1	2	3	4	5	6	7	$r_2 \gg 1$
1	1.1202	1.4123×10^{-1}	-2.7189×10^{-1}	-3.9227×10^{-1}	-4.2274×10^{-1}	-4.3027×10^{-1}	-4.3213×10^{-1}	-4.3259×10^{-1}	-4.3275×10^{1}
2	1.0004	-8.4580×10^{-1}	-1.7700	-1.9371	-1.9608	-1.9651	-1.9660	-1.9662	-1.9662
3	1.0000	-9.5391×10^{-1}	-1.9335	-1.9957	-1.9978	-1.9979	-1.9980	-1.9980	-1.9980
4	1.0000	-9.8514×10^{-1}	-1.9788	-1.9996	-1.9998	-1.9998	-1.9998	-1.9998	-1.9998
5	1.0000	-9.9475×10^{-1}	-1.9529	-1.9396	-1.9358	-1.9350	-1.9348	-1.9348	-1.9347
6	1.0000	-9.0946×10^{-1}	1.2255	1.5740	1.6255	1.6364	1.6390	1.6396	1.6398

Table C.21: *The numerical orders of convergence to zero with respect to N_x when $N_x = N_y$ for R_3, denoted by α_3, using the Upwind scheme. The numerical results are generated using initial condition regularities $r_2 = 0, \ldots, 7$ and $r_2 \gg 1$, by considering $N_x = 3^\gamma$ for $\gamma = 1, \ldots, 7$ and fixed $L = 4$ and calculating them through α_3 as in Equation (C.2). The values of γ listed in the table are the smaller of the two values of γ used to generate the N_x required to calculate α_3. These values were generated using $\mu_1 = \mu_2 = 1$ and $h_1 = h_2 = \frac{1}{2}$. Equation (5.131) was used to generate the order of convergence when $r_2 \gg 1$.*

C.3.2 analytically for the Upwind scheme

The analytical order of convergence of R_3 for the Upwind scheme is calculated in the following.

$$\begin{aligned}
&\sum_{q=2}^{\frac{N_y-1}{2}} \frac{|1-\nu_{1,q}|^2}{|q-1|^{2(r_2+1)}}, \\
= \quad & \frac{K_{3a}^2 L^2}{N_y^4} \sum_{q=1}^{\frac{N_y-1}{2}} q^{2-2r_2}, \\
< \quad & \begin{cases} \frac{K_{3a}^2}{24} L^2 N_y^{-1}, & \text{for } r_2 = 0, \\ \frac{K_{3a}^2}{2} L^2 N_y^{-3}, & \text{for } r_2 = 1, \\ K_{3a}^2 \zeta(2r_2-2) L^2 N_y^{-4}, & \text{for } r_2 \geq 2, \end{cases}
\end{aligned}$$

$$\Rightarrow \quad E_3 = N_x N_y \sum_{q=2}^{\frac{N_y-1}{2}} \frac{|1-\nu_{1,q}|^2}{|q-1|^{2(r_2+1)}} < \begin{cases} \frac{K_{3a}^2}{24} L^2 N_x, & \text{for } r_2 = 0, \\ \frac{K_{3a}^2}{2} L^2 N_x N_y^{-2}, & \text{for } r_2 = 1, \\ K_{3a}^2 \zeta(2r_1-2) L^2 N_x N_y^{-3}, & \text{for } r_2 \geq 2, \end{cases} \tag{C.29}$$

$$\Rightarrow \quad 2N_x N_y A_2^2 \sum_{q=2}^{\frac{N_y-1}{2}} \frac{|1-\nu_{1,q}|^2}{|q-1|^{2(r_2+1)}} = \begin{cases} \mathcal{O}(L^2 N_x), & \text{for } r_2 = 0, \\ \mathcal{O}(L^2 N_x N_y^{-2}), & \text{for } r_2 = 1, \\ \mathcal{O}(L^2 N_x N_y^{-3}), & \text{for } r_2 \geq 2. \end{cases} \tag{C.30}$$

When $N_x = N_y$ this results in,

$$2N_x N_y A_2^2 \sum_{q=2}^{\frac{N_y-1}{2}} \frac{|1-\nu_{1,q}|^2}{|q-1|^{2(r_2+1)}} = \begin{cases} \mathcal{O}(L^2 N_x), & \text{for } r_2 = 0, \\ \mathcal{O}(L^2 N_x^{-1}), & \text{for } r_2 = 1, \\ \mathcal{O}(L^2 N_x^{-2}), & \text{for } r_2 \geq 2. \end{cases} \tag{C.31}$$

C.3.3 with respect to N_xN_y for the Crank-Nicolson scheme

Crank-Nicolson Scheme: $\alpha_3 = \log\left(\frac{E_{3,3N_x}}{E_{3,N_x}}\right)/\log(3)$									
γ \ r_2	0	1	2	3	4	5	6	7	$r_2 \gg 1$
1	6.3590×10^{-1}	-1.1610	-2.3905	-2.9133	-3.0578	-3.0925	-3.1008	-3.1029	-3.1035
2	1.0346	-9.8670×10^{-1}	-2.8564	-3.7582	-3.8994	-3.9174	-3.9206	-3.9212	-3.9214
3	1.0015	-9.9948×10^{-1}	-2.9574	-3.9261	-3.9873	-3.9904	-3.9908	-3.9909	-3.9909
4	9.9980×10^{-1}	-1.0001	-2.9863	-3.9760	-3.9985	-3.9989	-3.9990	-3.9990	-3.9990
5	1.0000	-1.0000	-2.9955	-3.9921	-3.9998	-3.9999	-3.9999	-3.9999	-3.9999
6	1.0000	-1.0000	-2.9985	-3.9974	-4.0000	-4.0000	-4.0000	-4.0000	-4.0000

Table C.22: *The numerical orders of convergence to zero with respect to N_x when $N_x = N_y$ for R_3, denoted by α_3, using the Crank-Nicolson scheme. The numerical results are generated using initial condition regularities $r_2 = 0, \ldots, 7$ and $r_2 \gg 1$, by considering $N_x = 3^\gamma$ for $\gamma = 1, \ldots, 7$ and fixed $L = 4$ and calculating them through α_3 as in Equation (C.2). The values of γ listed in the table are the smaller of the two values of γ used to generate the N_x required to calculate α_3. These values were generated using $\mu_1 = \mu_2 = 1$ and $h_1 = h_2 = \frac{1}{2}$. Equation (5.131) was used to generate the order of convergence when $r_2 \gg 1$.*

C.3.4 analytically for the Crank-Nicolson scheme

The analytical order of convergence of R_3 for the Upwind scheme is calculated in the following.

$$\begin{aligned}
&\sum_{q=2}^{\frac{N_y-1}{2}} \frac{|1-\nu_{1,q}|^2}{|q-1|^{2(r_2+1)}}, \\
= \quad & \frac{K_{4b}^2 L^2}{N_y^6} \sum_{q=1}^{\frac{N_y-1}{2}} p^{4-2r_2}, \\
< \quad & \begin{cases} \frac{K_{4b}^2}{160} L^2 N_y^{-1}, & \text{for } r_2 = 0, \\ \frac{K_{4b}^2}{6} L^2 N_y^{-3}, & \text{for } r_2 = 1, \\ \frac{K_{4b}^2}{2} L^2 N_y^{-5}, & \text{for } r_2 = 2, \\ K_{4b}^2 \zeta(2r_2-4) L^2 N_y^{-6}, & \text{for } r_2 \geq 3, \end{cases}
\end{aligned}$$

$$\Rightarrow \quad E_3 = N_x N_y \sum_{q=2}^{\frac{N_y-1}{2}} \frac{|1-\nu_{1,q}|^2}{|q-1|^{2(r_2+1)}} < \begin{cases} \frac{K_{4b}^2}{160} L^2 N_y, & \text{for } r_2 = 0, \\ \frac{K_{4b}^2}{6} L^2 N_x N_y^{-2}, & \text{for } r_2 = 1, \\ \frac{K_{4b}^2}{2} L^2 N_x N_y^{-4}, & \text{for } r_2 = 2, \\ K_{4b}^2 \zeta(2r_2-4) L^2 N_x N_y^{-5}, & \text{for } r_2 \geq 3, \end{cases} \tag{C.32}$$

$$\Rightarrow \quad 2N_x N_y A_2^2 \sum_{q=2}^{\frac{N_y-1}{2}} \frac{|1-\nu_{1,q}|^2}{|q-1|^{2(r_2+1)}} = \begin{cases} \mathcal{O}(L^2 N_x), & \text{for } r_2 = 0, \\ \mathcal{O}(L^2 N_x N_y^{-2}), & \text{for } r_2 = 1, \\ \mathcal{O}(L^2 N_x N_y^{-4}), & \text{for } r_2 = 2, \\ \mathcal{O}(L^2 N_x N_y^{-5}), & \text{for } r_2 \geq 3. \end{cases} \tag{C.33}$$

When $N_x = N_y$ this results in,

$$2N_x N_y A_2^2 \sum_{q=2}^{\frac{N_y-1}{2}} \frac{|1-\nu_{1,q}|^2}{|q-1|^{2(r_2+1)}} = \begin{cases} \mathcal{O}(L^2 N_x), & \text{for } r_2 = 0, \\ \mathcal{O}(L^2 N_x^{-1}), & \text{for } r_2 = 1, \\ \mathcal{O}(L^2 N_x^{-3}), & \text{for } r_2 = 2, \\ \mathcal{O}(L^2 N_x^{-4}), & \text{for } r_2 \geq 3. \end{cases} \tag{C.34}$$

BIBLIOGRAPHY

[1] R. Daley. Estimating model-error covariances for application to atmospheric data assimilation. *Monthly Weather Review*, 120(8):1735–1746, 1992.

[2] A. C. Lorenc. Analysis methods for numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, 112(474):1177–1194, 1986.

[3] F. X. Le Dimet and V. Shutyaev. On deterministic error analysis in variational data assimilation. *Nonlinear Processes in Geophysics*, 12(4):481–490, 2005.

[4] F. Rabier, H. Järvinen, E. Klinker, J. F. Mahfouf, and A. Simmons. The ECMWF operational implementation of four-dimensional variational assimilation. I: experimental results with simplified physics. *Quarterly Journal of the Royal Meterological Society*, 126(564):1143–1170, 2000.

[5] S. Haben. *Conditioning and Preconditioning of the Minimisation Problem in Variational Data Assimilation.* Thesis, PhD in Mathematics, University of Reading, 2011.

[6] D. R. Durran. *Numerical Methods for Wave Equations in Geophysical Fluid Dynamics*, volume 32 of *Texts in Applied Mathematics Series.* New York: Springer New York, 1999.

[7] L. N. Trefethen. Finite difference and spectral methods for ordinary and partial differential equations. available from: "`http://people.maths.ox.ac.uk/trefethen/pdetext.html`" [accessed 09/05/2011], 1996.

[8] H. O. Kreiss. Initial boundary value problems for hyperbolic systems. *Communications on Pure and Applied Mathematics*, 23(3):277–298, 1970.

[9] E. F. Toro. *Shock-Capturing Methods for Free-Surface Shallow Flows.* John Wiley & Sons, Ltd., 2001.

[10] E. V. Hólm. Lecture notes on assimilation algorithms. *Meteorological Training Course Lecture Series*, 2003. available from: "`http://old.ecmwf.int/newsevents/training/lecture_notes/pdf_files/ASSIM/Ass_algs.pdf`" [accessed: 24/07/2010].

[11] A. S. Lawless. Variational data assimilation for very large environmental problems. In M. Cullen, M. A. Freitag, S. Kindermann, and R. Scheichl, editors, *Large Scale Inverse Problems. Computational Methods and Applications in the Earth Sciences*, volume 13 of *Radon Series on Computional and Applied Mathematics*, pages 55–90. Walter de Gruyter, Berlin, 2013.

[12] P. Courtier, Thépaut J. N., and A. Hollingsworth. A strategy for operational implementation of 4d-var, using an incremental approach. *Quarterly Journal of the Royal Meteorological Society*, 120B(519):1367–1387, 1994.

[13] D. M. Causon and C. G. Mingham. *Introductory Finite Difference Methods for PDEs*. bookboon.com [accessed 08/02/2014], 2010.

[14] K. W. Morton and D. F. Mayers. *Numerical Solution of Partial Differential Equations*. Cambridge University Press, Cambridge, UK, 2nd edition, 2005.

[15] R. Vichnevetsky and J. B. Bowles. *Fourier Analysis of Numerical Approximations of Hyperbolic Equations*. SIAM, 1982.

[16] M. A. Freitag and R. W.E. Potthast. Synergy of inverse problems and data assimilation techniques. In M. Cullen, M. A. Freitag, S. Kindermann, and R. Scheichl, editors, *Large Scale Inverse Problems. Computational Methods and Applications in the Earth Sciences*, volume 13 of *Radon Series on Computational and Applied Mathematics*, pages 1–54. Walter de Gruyter, Berlin, 2013.

[17] C. Johnson, N. K. Nichols, and B. J. Hoskins. Very large inverse problems in atmosphere and ocean modelling. *International Journal for Numerical Methods in Fluids*, 47(8-9):759–771, 2005.

[18] F. Bouttier and P. Courtier. Data assimilation concepts and methods. *Meteorological Training Course Lecture Series*, 1999. available from: "`http://msi.ttu.ee/~elken/Assim_concepts.pdf`" [accessed: 24/07/2010].

[19] A. F. Bennet. *Inverse Methods in Physical Oceanography*. Cambridge University Press, 1992.

[20] R. Daley. *Atmospheric Data Analysis*. Cambridge Atmospheric and Space Science Series. Cambridge: Cambridge University Press, 1999.

[21] N. K. Nichols. Mathematical concepts of data assimilation. In W. Lahoz, B. Khattatov, and B. Menard, editors, *Data Assimilation, Making Sense of Observations*, pages 13–39. Springer Berlin Heidelberg, 2010.

[22] C. K. Wikle and L. M. Berliner. A bayesian tutorial for data assimilation. *Physica D: Nonlinear Phenomena*, 230(1):1–16, 2007.

[23] Met. Office. Met. Office unified model. *Met. Office*, 2014. available from: "`http://www.metoffice.gov.uk/research/modelling-systems/unified-model`" [accessed 07/10/2014].

[24] Met. Office. Met. Office numerical weather prediction models. *Met. Office*, 2014. available from: "`http://www.metoffice.gov.uk/research/modelling-systems/unified-model/weather-forecasting`" [accessed 07/10/2014].

[25] R. Buizza. Chaos and weather prediction. *Meteorological Training Course Lecture Series, ECMWF*, 2000. available from: "`http://old.ecmwf.int/newsevents/training/lecture_notes/pdf_files/PREDICT/Chaos.pdf`" [accessed: 24/07/2010].

[26] F. Möller and E. Raschke. Problems of meteorological observations from satellites. *Space Science Reviews*, 9(1):90–148, 1969.

[27] D. P. Dee and A. M. Da Silva. Data assimilation in the presence of forecast bias. *Quarterly Journal of the Royal Meteorological Society*, 124A(545):269–295, 1998.

[28] E. E. Petrie. *Background error covariance modelling for convective-scale variational data assimilation.* Thesis, PhD, University of Reading, 2012.

[29] M. Bonavita and E. Isaksen, L.Hólm. On the use of eda background error covariances in the ecmwf 4d-var. *Quarterly Journal of the Royal Meteorology Society*, 138(667):1540–1559, 2012.

[30] M. J. P. Cullen. Four-dimensional variational data assimilation: A new formulation of the background-error covariance matrix based on a potential-vorticity representation. *Quarterly Journal of the Royal Meteorology Society*, 129(593):2777–2796, 2003.

[31] M. J. P. Cullen. A demonstration of 4d-var using a time-distributed background term. *Quarterly Journal of the Royal Meteorology Society*, 136(650):1301–1315, 2010 Part A.

[32] G. A. Hajj, E. R. Kursinski, L. J. Romans, E. I. Bertiger, and S. S. Leroy. A technical description of atmospheric sounding by gps occultation. *Journal of Atmospheric and Solar-Terrestrial Physics*, 64(4):451–469, 2002.

[33] J. N. Thépaut, R. N. Hoffman, and P. Courtier. Interactions of dynamics and observations in a four-dimensional variational assimilation. *Monthly Weather Review*, 121(12):3393–3414, 1993.

[34] E. Simon and L. Bertino. Gaussian anamorphosis extension of the DEnKF for combined state parameter estimation: application to a 1D ocean ecosystem model. *Journal of Marine Systems*, 89(1):1–18, 2012.

[35] F. X. Le Dimet and O. Talagrand. Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. *Tellus A*, 38(2):97–110, 1986.

[36] C. Johnson, B. J. Hoskins, and N. K. Nichols. A singular vector perspective of 4D-Var: filtering and interpolation. *Quaterly Journal of the Royal Meteorological Society*, 131(605):1–19, 2005.

[37] A. S. Lawless, S. Gratton, and N. K. Nichols. An investigation of incremental 4d-var using non-tangent linear models. *Quarterly Journal of the Royal Meteorological Society*, 131(606):459–476, 2005.

[38] J. M. Lewis, S. Lakshmivarahan, and S. K. Dhall. *Dynamic Data Assimilation: A Least Squares Approach*. Encyclopedia of Mathematics and its Applications: 104. Cambridge University Press, 2006.

[39] A. S. Lawless, N. K. Nichols, and S. P. Ballard. A comparison of two methods for developing the linearization of a shallow-water model. *Quarterly Journal of the Royal Meterological Society*, 129(589):1237–1254, 2003.

[40] F. Rawlins, S. P. Ballard, K. J. Bovis, A. M. Clayton, D. Li, G. W. Inverarity, A. C. Lorenc, and T. J. Payne. The Met. Office global four-dimensional variational data assimilation scheme. *Quarterly Journal of the Royal Meteorological Society*, 133(623):347–362, 2007.

[41] M. A. Freitag, N. K. Nichols, and C. J. Budd. Resolution of sharp fronts in the presence of model error in variational data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 139(672A):742–757, 2013.

[42] X. Zou, I. M. Navon, and F. X. Le Dimet. Incomplete observations and control of gravity waves in variational data assimilation. *Tellus A*, 44(4):273–296, 1992.

[43] R. Ménard and R. Daley. The application of kalman smoother theory to the estimation of 4d-var error statistics. *Tellus A*, 48(2):221–227, 1996.

[44] Y. Sasaki. Some basic formalisms in numerical variational analysis. *Monthly Weather Review*, 98(12):875–883, 1970.

[45] A. K. Griffith and N. K. Nichols. Adjoint methods in data assimilation for estimating model error. *Flow, Turbulence and Combustion*, 65(3–4):469–488, 2000.

[46] P. A. Vidard, A. Piacentini, and F. X. Le Dimet. Variational data analysis with control of the forecast bias. *Tellus A*, 56(3):177–188, 2004.

[47] S. Akella and I. M. Navon. Different approaches to model error formulation in 4D-Var: a study with high-resolution advection schemes. *Tellus A*, 61:112–128, 2009.

[48] J. C. Derber. A variational continuous assimilation technique. *Monthly Weather Review*, 117(11):2437–2446, 1989.

[49] M. Zupanski. Regional four-dimensional variational data assimilation in quasi-operational forecasting environment. *Monthly Weather Review*, 121(8):2396–2408, 1993.

[50] W. Wergen. The effect of model errors in variational assimilation. *Tellus A*, 44(4):297–313, 1992.

[51] P. A. Vidard, E. Blayo, F. X. Le Dimet, and A. Piacentini. 4D variational data analysis with imperfect model. *Flow, Turbulence and Combustion*, 65(3–4):489–504, 2000.

[52] D. Furbish, M. Y. Hussaini, F. X. Le Dimet, P. Ngnepieba, and Y. Wu. On discretization error and its control in variational data assimilation. *Tellus A*, 60(5):979–991, 2008.

[53] Y. Trémolet. Model-error estimation in 4D-Var. *Quarterly Journal of the Royal Meteorological Society*, 133(626):1267–1280, 2007.

[54] R. Gerdes, C. Köberle, and J. Willebrand. The influence of numerical advection schemes on the results of ocean general circulation models. *Climate Dynamics*, 5(4):211–226, 1991.

[55] T. Vukićević, M. Steyskal, and M. Hecht. Properties of advection algorithms in the context of variational data assimilation. *Monthly Weather Review*, 129(5):1221–1231, 2001.

[56] E. A. Celaya and J. J. Anza. BDF-α: A multistep method with numerical damping control. *Universal Journal of Computational Mathematics*, 1(3):96–108, 2013.

[57] B. M. Broderick, A. S. Elnashai, and B. A. Izzuddin. Observations on the effect of numerical dissipation on the nonlinear dynamic response of structural systems. *Engineering Structures*, 16(1):51–62, 1994.

[58] R. B. Rood. Numerical advection algorithms and their role in atmospheric transport and chemistry models. *Reviews of Geophysics*, 25(1):71–100, 1987.

[59] R. J. Le Veque. *Numerical Methods for Conservation Laws*. Lectures in Mathematics. Basel: Birkhäuser Verlag, 1999.

[60] W. L. Briggs and V. E. Henson. *The DFT : an owner's manual for the discrete Fourier transform*. SIAM, 1995.

[61] H. S. Carslaw. *Introduction to the Theory of Fourier's Series and Integrals*. New York: Dover Publications, Inc., 3rd edition, 1950.

[62] M. R. Spiegel. *Theory and Problems of Fourier Analysis with Applications to Boundary Value Problems*. Schaum's Outlines. McGraw-Hill, 1974.

[63] R. V. Churchill and J. W. Brown. *Fourier Series and Boundary Value Problems*. McGraw-Hill, Inc., 3rd edition, 1941.

[64] M. A. Freitag. Transcritical flow modelling with the box scheme. Dissertation, MSc in modern applications of mathematics, University of Bath, 2003.

[65] P. C. Hansen, J. G. Nagy, and D. P. O'Leary. *Deblurring Images: Matrices, Spectra, and Filtering*. Fundamentals of Algorithms. SIAM, 2006.

[66] V. E. Henson. *Fourier Methods of Image Reconstruction*. Thesis, PhD in Mathematics, University of Colorado, 1990.

[67] R. W. Hamming. *Numerical Methods for Scientists and Engineers*. McGraw-Hill, 2nd edition, 1915.

[68] P. D. Williams. A proposed modification to the robert-asselin time filter. *Monthly Weather Review*, 137(8):2538–2546, 2009.

[69] C. .B. Vreugdenhil. *Numerical Methods for Shallow-Water Flow*, volume 13 of *Water Science and Technology Library*. Kluwer Academic Publishers, 1994.

[70] R. J. Le Veque. *Finite Volume Methods for Hyperbolic Problems*. Cambridge Texts in Applied Mathematics. Cambridge: Cambridge University Press, 2002.

[71] J. C. Strikwerda. *Finite Difference Schemes and Partial Differential Equations*. The Wadsworth & Brooks/Cole Mathematics Series. Wadsworth & Brooks/Cole Advanced Books & Software, 1989.

[72] E. F. Toro. *Riemann Solvers and Numerical Methods for Fluid Dynamics*. Springer, 3rd edition, 2009.

[73] R. Courant, K. Freidrichs, and H. Lewy. On the partial difference equations of mathematical physics. *IBM Journal of Research and Development*, 11(2):215–234, 1967.

[74] Mathworks, Inc. 1994-2013. *MATLAB®*, 2012b edition, 2012.

[75] R. J. Beerends, H. G. ter Morsche, J. C. van den Berg, and E. M. van de Vrie. *Fourier and Laplace Transforms*. Cambridge: Cambridge University Press, 2003.

[76] J. P. Boyd. *Chebyshev and Fourier Spectral Methods*. Dover Publications, Inc., 2nd (revised) edition, 2001.

[77] T. M. Apostol. *Mathematical Analysis*. Addison-Wesley Publishing Company, 2nd edition, 1974.

[78] J. B. Martens. The hermite transform-theory. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 38(9):1595–1606, 1990.

[79] S. K. Mitra. *Digital Signal Processing: A Computer Based Approach*. McGraw Hill Higher Education, 3rd edition, 2006.

[80] G. R. Grimmett and D. R. Stirzaker. *Probability and random processes*. Oxford: Clarendon Press, 1982.

[81] A. K. Griffith. *Data Assimilation for Numerical Weather Prediction Using Control Theory*. Thesis, PhD in Mathematics, University of Reading, 1997.

[82] A. K. Griffith and N. K. Nichols. *Data Assimilation Using Optimal Control Theory*. Numerical analysis report, University of Reading, 1994.

[83] R. L. Pfeffer, I. M. Navon, and X. Zou. A comparison of the impact of two time-differencing schemes on the NASA-GLAS climate model. *Monthly Weather Review*, 120(7):1381–1393, 1992.

[84] J. Stewart. *Multivariable Calculus*. Brooks/Cole, Cengage Learning, 7th international edition, 2012. Metric Version.

[85] W. H. Young. On multiple integration by parts and the second theorem of the mean. *Proceedings of the London Mathematical Society*, s2-16(1):273–293, 1917.

[86] R. Courant. *Differential and Integral Calculus*, volume I. Blackie & Son Limited, 2nd edition, 1964.

[87] M. J.P. Cullen. *A Mathemtical Theory of Large-Scale Atmosphere/Ocean Flow*. London: Imperial College Press, 2006.

[88] A. I. Delis and Th. Katsaounis. Relaxation schemes for the shallow water equations. *International Journal for Numerical Methods in Fluids*, 41(7):695–719, 2003.

[89] T. Chacón Rebollo, A. Domínguez Delgado, and E. D. Fernández Nieto. A family of stable numerical solvers for the shallow water equations with source terms. *Computer Methods in Applied Mechanics and Engineering*, 192(1):203–225, 2003.

[90] J. H. Kwak and S. Hong. *Linear Algebra*. Birkhäuser Boston, 2nd edition, 2004.

[91] D. R. Durran. *Numerical Methods for Fluid Dynamics with Applications to Geophysics*. New York: Springer New York, 2010.

[92] J. R. Cardoso and F. Silva Leite. Exponentials of skew-symmetric matrices and logarithms of orthogonal matrices. *Journal of Computational and Applied Mathematics*, 233(11):2867–2875, 2010.

[93] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.

[94] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical Mathematics*. Texts in Applied Mathematics, 37. Springer, 2nd edition, 2006.

[95] D. Serre. *Matrices: theory and applications*. Graduate texts in mathematics: 216. Springer, 2002.

[96] A. Ibrahimbegovic. *Nonlinear Solid Mechanics*. Solid Mechanics and its Applications: 160. Dordrecht: Springer Netherlands, 2009.

[97] E. W. Weisstein. "power sum". *From MathWorld–A Wolfram Resource*, 2012. available from: "`http://mathworld.wolfram.com/PowerSum.html`" [accessed 24/06/2012].

[98] J. Sondow and E. W. Weisstein. "harmonic number.". *From MathWorld–A Wolfram Web Resource*, 2014. available from: "`http://mathworld.wolfram.com/HarmonicNumber.html`" [accessed 11/07/2014].